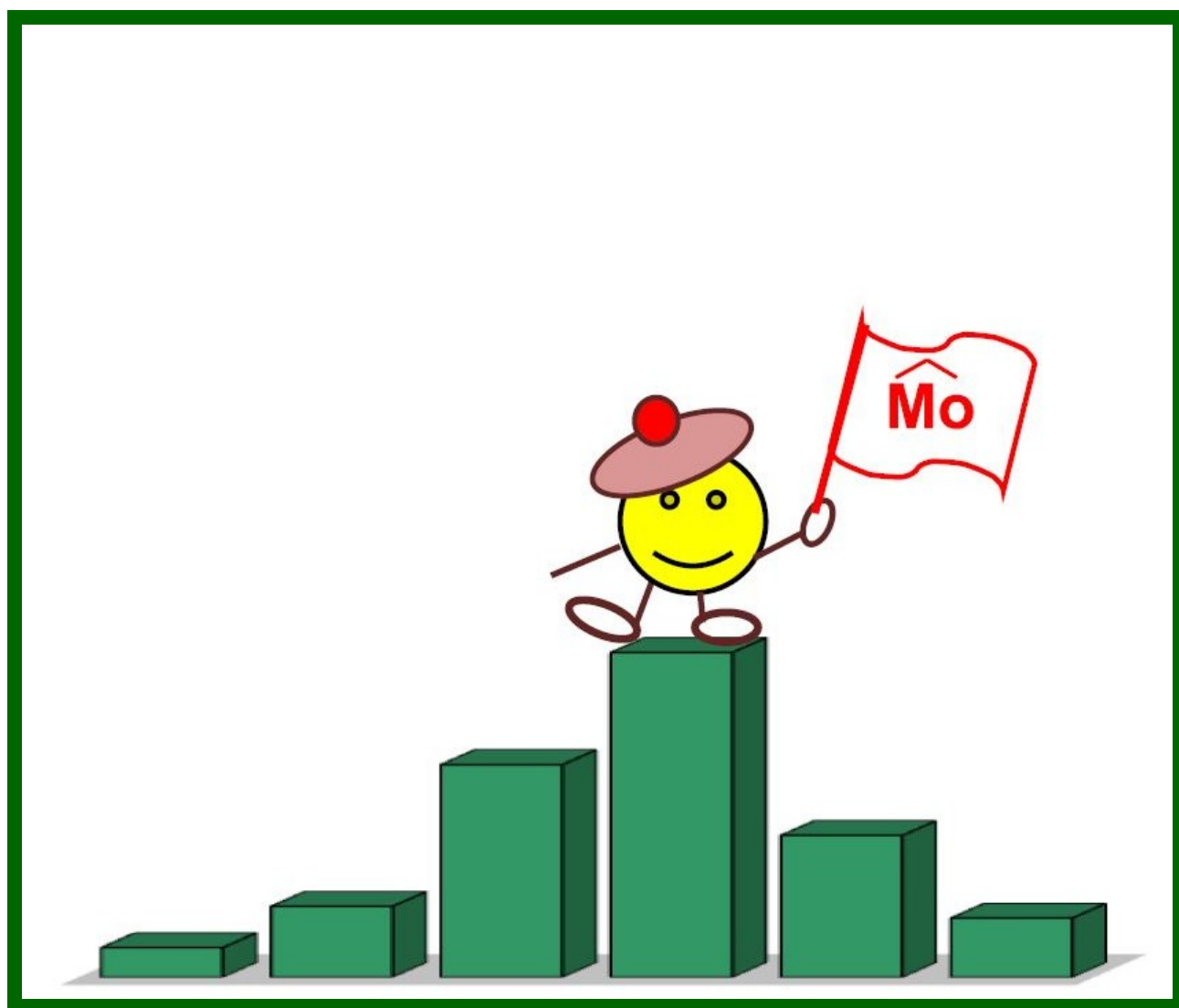


Е.К. Каштанова

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА



Казань – 2016

УДК 519.2(07)

ББК 22.172я7

К31

Печатается по решению Редакционно-издательского совета ФГАОУВПО

«Казанский (Приволжский) федеральный университет»

методической комиссии ИВМиИТ

Протокол № 9 от 12 мая 2016 г.

заседания кафедры математической статистики

Протокол № 9 от 10 мая 2016 г.

Научный редактор

канд. физ.мат. наук С.Г.Халиуллин

Рецензенты

доктор физ.-мат. наук И.Н.Володин

доктор физ.-мат. наук В.С.Желтухин

Дизайн обложки Е.К.Каштанова

Каштанова Е.К.

К31 Математическая статистика: учебное пособие/ Е.К.Каштанова – Казань: Изд-во Казан. ун-та, 2016. – 196 с.

ISBN 978-5-00019-741-7

Учебное пособие «Математическая статистика» предназначено для обучения основам математической статистики студентов социальных и экономических специальностей. При разработке курса большое внимание было уделено практической направленности курса. Учебное пособие «Математическая статистика» предназначено для развития профессиональной компетенции «умение обрабатывать и анализировать данные для подготовки аналитических решений, экспертных заключений и рекомендаций». Учебное пособие содержит теоретический материал, разбор типовых задач, задачи с ответами, таблицы, глоссарий, рекомендуемую литературу и Интернет-источники. Учебное пособие может быть рекомендовано к использованию преподавателями данной дисциплины и студентами в течение семестра и при подготовке к зачету и экзамену.

© Каштанова Е.К., 2016

© Издательство Казанского университета

СОДЕРЖАНИЕ

Введение.	5
1. Классификация признаков по уровням измерений.	7
2. Описательная статистика	12
2.1. Вариационный ряд и его графическое изображение. . .	12
2.2. Числовые характеристики выборки.	23
2.2.1. Характеристики положения	23
2.2.2. Характеристики вариации	26
2.2.3. Упрощенные методы расчета выборочной средней и выборочной дисперсии	28
3. Выборочный метод	40
3.1. Статистические оценки параметров распределения. . .	40
3.1.1. Точечные оценки	41
3.1.2. Интервальные оценки.	44
4. Проверка статистических гипотез	56
4.1. Проверка гипотезы о величине среднего значения нормального распределения	60
4.2. Проверка гипотезы о равенстве средних значений двух нормальных распределений	65
4.3. Проверка гипотезы о равенстве дисперсий двух нор- мальных распределений.	71
4.4. Проверка гипотезы о равенстве долей двух распреде- лений.	78
4.5. Критерий χ^2	85
5. Статистическое изучение связи	99
5.1. Корреляционный анализ	102
5.1.1. Линейная корреляция. Элементы регрессионного анализа	105
5.1.2. Ранговая корреляция	125
5.1.2.1. Коэффициент ранговой корреляции Спирмена . .	127
5.1.2.2. Коэффициент ранговой корреляции Кендалла. . .	131

5.1.2.3. Коэффициент конкордации	139
5.2. Ранговый бисериальный коэффициент.	144
5.3. Точечно-бисериальный коэффициент.	146
5.4. Коэффициенты связи, основанные на критерии χ^2 . . .	152
5.5. Коэффициенты ассоциации и контингенции.	156
Приложение	165
Глоссарий	180
Информационные источники	183
Ответы	188

Введение

Математическая статистика – раздел математики, изучающий математические методы сбора, систематизации, обработки и интерпретации результатов наблюдений с целью выявления статистических закономерностей.

Слово «статистика» в его современном значении впервые употребил немецкий ученый Готфрид Ахенваль (1719 – 1772), заимствовав его из итальянского языка. В Италии в период эпохи Возрождения получили распространение знания о политике, составившие специальную дисциплину, которая называлась *ragione di stato* или *disciplina de statu*. Слова *statu* и *statu* соответствовали понятию «государство» (отсюда немецкое *staat* и английское *state*).

В Оксфордском словаре английского языка статистика определяется как «собранные и классифицированные числовые данные и сведения», а также как «наука, изучающая методы сбора и обработки числовых данных, относящихся к человеческой деятельности и природным явлениям».

Термин «статистика» употребляется в нескольких значениях: 1) количественные данные; 2) параметр совокупности; 3) наука о массовых явлениях; 4) универсальный метод научного познания.

В настоящее время статистические методы применяют практически все – от политиков, желающих предсказать исход выборов, до предпринимателей, стремящихся оптимизировать прибыль. Такому широкому распространению методов математической статистики способствовало развитие вычислительной техники, особенно персональных компьютеров. На сегодняшний день создано множество универсальных и специализированных статистических пакетов. Перечислим только наиболее популярные в России: отечественные STADIA, Эвриста, Статистик-консультант, Олимп: СтатЭксперт, Мезозавр, Класс-Мастер и американские STATGRAPHICS, SAS, SPSS, SYSTAT, STATISTICA и др. В самом распространенном пакете – Ex-

cel – можно выполнить практически все расчеты, построить разнообразные графики и диаграммы. Компьютер в наше время может практически все. Но чтобы получить действительно правильные результаты, нужно разбираться в методах математической статистики, знать условия их применения, грамотно ставить задачу, правильно делать выводы.

Целью любого статистического исследования является принятие решения относительно обследуемого объекта. Такое решение принимается с помощью специальных статистических правил на основе наблюдения некоторых характеристик объекта (обычно числовых или векторных). Результаты наблюдений трактуются как реализации случайных величин, и поэтому возможны ошибки в принятии решений. Математическая статистика занимается построением статистических правил, минимизирующих частоту принятия неправильных решений, а также вычислением этой частоты для традиционных (известных) статистических правил.

Все выводы математической статистики основываются на статистических данных. Приведем наиболее простой и распространенный способ получения данных.

Пусть X – наблюдаемая характеристика (случайная величина – с.в.) исследуемого объекта. Проведем серию опытов, состоящую из n раз независимых наблюдений X . В результате мы получаем n наблюдений, которые обозначим $x_1, x_2, \dots, x_n$. Обычно в математической статистике они называются выборкой из распределения случайной величины X (устаревшее – «выборка из генеральной совокупности значений X ») и представляют собой, строго говоря, результат наблюдения n независимых копий $X_1, \dots, X_n$ с.в. X .

Следующие задачи являются наиболее типичными для математической статистики.

1. Оценка неизвестных параметров в распределении наблюдаемых характеристик (случайных величин),
2. Проверка гипотез (высказываний, предположений) о распределении наблюдаемых случайных величин.

1. КЛАССИФИКАЦИЯ ПРИЗНАКОВ ПО УРОВНЯМ ИЗМЕРЕНИЙ

Ключевые слова: количественные признаки, качественные признаки, измерение, шкала, номинальная шкала, порядковая шкала, интервальная шкала, шкала отношений.

Экспериментальные данные обычно представляют собой результаты измерения некоторых признаков. Признаки делятся на качественные и количественные.

Количественные признаки представляют собой результаты подсчета и измерения (например, трудовой стаж, число детей в семье, заработная плата, рост и вес человека и т. п.).

Качественные признаки – это признаки, которыми объект либо обладает, либо не обладает. Они не поддаются непосредственному измерению (например, пол человека, национальность, квалификация и т.д.). Чтобы к качественным признакам можно было применять математические методы, их кодируют в числовой форме.

Измерение – это приписывание числовых форм объектам или событиям в соответствии с определенными правилами.

Шкала – правило, по которому в процессе измерения каждому изучаемому объекту ставится в соответствие некоторое число

В настоящее время существует очень большое число типов измерительных шкал, но обычно выделяют 4 основных типа:

- номинальная,
- порядковая,
- интервальная,
- шкала отношений.

Номинальная и порядковая шкалы используются в большей степени для качественных признаков, а интервальная шкала и шкала отношений – для количественных признаков. Такое деление на шкалы является весьма условным, потому что один и тот же признак можно представить в виде разных шкал. Все зависит от задач исследования. Однако следует учитывать, что неправильная идентификация признака может привести к искаженному результату.

Номинальная шкала

Номинальная (номинативная или шкала наименований) – это шкала, классифицирующая по принципу «равенства-неравенства» между изучаемыми объектами.

Измерение объекта по номинальной шкале фактически сводится к классификации объекта в один из нескольких непересекающихся классов. При этом объекты, принадлежащие одному классу, считаются по данному признаку идентичными, а к разным – различными.

Пример 1. Признак – пол человека.

1. Мужчина
2. Женщина

Пример 2. Признак – профессия человека.

1. Токарь
2. Слесарь
3. Монтажник
4. Фрезеровщик

При обработке номинальных данных очень часто каждому объекту ставится в соответствие символ или число (например, номер класса). Такая кодировка очень удобна для обработки данных на компьютере, но следует учитывать, что такие данные только формально являются количественными, а по своей сути остаются качественными.

Порядковая шкала

Порядковая (ранговая, ординальная) – это шкала, классифицирующая по принципу «больше – меньше», «выше – ниже» и т. п.

В порядковой шкале все признаки располагаются по степени нарастания, увеличения признака, т.е. по рангу – от самого большого до самого маленького. Или наоборот.

Пример 1. Варианты ответов в социологических опросах: «Часто/иногда/никогда».

Пример 2. Турнирная таблица спортивных соревнований.

Результаты в беге на 100 м

Место	Спортсмен	Результат
1	А.	12,8
2	Б.	13,0
3	В.	15,8
4	Г.	15,9

Спортсмен, пришедший 2-м, отстал от первого спортсмена всего лишь на 0,2 с. В то время как 3-й спортсмен отстал от 2-го на 2,8 с. Т.е. мы не можем сказать, что различие между 1-м и 2-м спортсменами так же велико, как между 2-м и 3-м, между 3-м и 4-м. Мы можем лишь утверждать, что быстрее всех прибежал спортсмен А., остальные достигли финиша в следующей последовательности: Б., В., Г.

Интервальная шкала

Интервальная шкала – это шкала, классифицирующая по принципу «больше на определенное количество единиц – меньше на опре-

деленное количество единиц». Существует единица измерения. Расположение точки начала отсчета – произвольное.

Подобные шкалы используются для измерения пространственных и временных положений. В социологии нередко используются псевдоинтервальные шкалы (шкала Терстоуна, «термометр общественного мнения и т. п.).

Пример. Почти все температурные шкалы (Цельсия, Реомюра, Фаренгейта) имеют произвольную точку начала отсчета – нуля. Так, в шкале Цельсия за точку отсчета взята температура замерзания воды. Исключением является шкала Кельвина. Во всех шкалах произволен и масштаб измерения температуры. Например, между шкалами Фаренгейта (X') и Цельсия (X) существует следующее соответствие: $X' = 32 + 1,8X$.

Шкала отношений

Шкала отношений – это шкала, классифицирующая объекты или субъекты пропорционально степени выраженности измеряемого свойства.

Шкала отношений отличается от интервальной наличием точно зафиксированной точки начала отсчета.

Пример. Возраст человека исчисляется с момента рождения. Общепринятая единица измерения – 1 год.

Контрольные вопросы

1. Что является предметом изучения математической статистики?
2. Перечислите типичные задачи для математической статистики.
3. Что такое шкала?
4. В чем различие между шкалой порядка и интервальной шкалой?
5. Приведите пример порядковой шкалы.

6. Приведите пример интервальной шкалы.

Задачи к §1

Определите, к какому типу шкал можно отнести следующие данные.

1. «Скажите, пожалуйста, как часто вы используете Интернет не для работы или учёбы, а для себя, для собственных нужд, интереса?»

Варианты ответов:

- несколько раз в день,
- один раз в день,
- несколько раз в неделю,
- один раз в неделю,
- несколько раз в месяц,
- один раз в месяц,
- реже, чем раз в месяц,
- я не использую интернет для собственных нужд,
- затрудняюсь ответить.

2. Данные: продолжительность использования Интернета.

3. Данные: возраст респондента.

4. «С какой целью Вы хотите получить образование по программе МВА?»

Варианты ответов:

- а) для карьерного роста;
- б) для более успешной работы на своей позиции;
- в) для открытия собственного бизнеса.

5. Данные: расходы на транспорт в месяц.

6. Данные: тарифный разряд.

7. Данные: состав домохозяйства.

2. ОПИСАТЕЛЬНАЯ СТАТИСТИКА

2.1. Вариационный ряд и его графическое изображение

Ключевые слова: вариационный ряд, дискретный вариационный ряд, интервальный вариационный ряд, частота, частость, накопленная частота, полигон, гистограмма, кумулята.

Случайные величины выборки $x_1, x_2, \dots, x_n$, расположенные в порядке возрастания их значений: $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$, образуют *вариационный ряд*. В частности, $x_{(1)} = \min\{x_1, \dots, x_n\}$, $x_{(n)} = \max\{x_1, \dots, x_n\}$. Элементы вариационного ряда $x_{(i)}$ называются *вариантами*, $i=1..n$.

Если в выборке $x_1, x_2, \dots, x_n$ значения x_i встречается n_i раз, то число n_i называется *частотой* элемента выборки x_i . Общая сумма частот будет равна объему выборки, т.е. $\sum n_i = n$.

Относительной частотой (частостью) называется отношение частоты к объему выборки

$$w_i = \frac{n_i}{n}.$$

Относительная частота показывает, какую долю занимает данный признак (например, значение x_1) в выборке.

*Накопленная частота n_i^** – это частота появления в рассматриваемой совокупности всех значений признака, которые меньше или равны данному, т.е. для которых $X \leq x_i$.

Накопленная частота представляет собой сумму частот данной варианты и всех предшествующих вариантов.

*Относительная накопленная частота w_i^** показывает долю всех значений признака, для которых $X \leq x_i$, и равна $w_i^* = \frac{n_i^*}{n}$.

Вариационный ряд даже для малых выборок ($n < 30$) выглядит громоздко и не очень удобен для анализа. Поэтому для большей информативности данные группируют.

Дискретным вариационным рядом называется перечень вариантов и соответствующих им частот или относительных частот (таблица 1).

Таблица 1

X	x_1	x_2	...	x_r
n_i	n_1	n_2	...	n_r

Если признаки варьируют в широком диапазоне, т.е. число различающихся вариантов велико, то данные целесообразно группировать в интервалы.

Интервальным вариационным рядом называется соответствие между интервалами и частотами, которые равны сумме частот вариантов, относящихся к интервалу (таблица 2).

Таблица 2

Границы интервалов	$(a_0 - a_1)$	$(a_1 - a_2)$	...	$(a_{m-1} - a_m)$
n_i	n_1	n_2	...	n_m

В интервальных вариационных рядах вместо частот могут быть использованы и относительные частоты.

Для повышения наглядности эмпирических распределений используется их графическое представление. Наиболее распространенными графиками являются гистограмма, полигон и кумулята.

Гистограмма представляет собой фигуру, состоящую из прямоугольников, высота которых пропорциональна числу наблюдений, попавших в данные интервалы числовой прямой.

Для гистограммы частот высота равна $h_i = \frac{n_i}{\Delta}$, где Δ – ширина интервала, а для гистограммы относительных частот – $h_i = \frac{w_i}{\Delta} = \frac{n_i}{n\Delta}$.

Алгоритм построения гистограммы.

1. Ширина интервала выбирается исходя из смысла данных, логики исследования. Возможно использование неравных интервалов.

Если же никаких отдельных требований к организации данных нет, то действуем в следующем порядке.

а) Число интервалов разбиения вычисляется по формуле Стерджесса:
 $r = 1 + 3,322 \lg(n)$.

б) Ширина интервала определяется по следующей формуле:

$$\Delta \approx \frac{x_{\max} - x_{\min}}{r - 1}.$$

в) Область значений наблюдаемой с.в. X разбивается на интервалы одинаковой длины Δ : (a_{i-1}, a_i) $a_i = a_{i-1} + \Delta$, $i = 1..r$. Причем, $a_0 = x_{\min} - \frac{\Delta}{2}$.

2. Вычисляются значения n_i – количество данных, попавших в i -й интервал (a_{i-1}, a_i) , $i = 1..r$.

3. Над каждым i -м интервалом строится прямоугольник высоты h_i , причем

а) для гистограммы частот $h_i = \frac{n_i}{\Delta}$;

б) для гистограммы относительных частот $h_i = \frac{n_i}{n\Delta}$.

Для гистограммы относительных частот площадь каждого i -го прямоугольника будет равна $h_i \cdot \Delta = \frac{n_i}{n\Delta} \cdot \Delta = \frac{n_i}{n}$, т.е. относительной частоте. Тогда площадь (при больших n) будет оценкой вероятности попадания с.в. X в i -й интервал. Площадь ступенчатой фигуры под графиком гистограммы будет равна единице. Таким образом, гистограмма

является статистическим аналогом функции плотности для непрерывных распределений.

Полигоном называется ломаная кривая, соединяющая точки $(x_1, h_1), (x_2, h_2), \dots, (x_r, h_r)$. Как и в случае гистограммы, для полигона частот высота равна $h_i = \frac{n_i}{\Delta}$, а для полигона относительных частот –

$$h_i = \frac{w_i}{\Delta} = \frac{n_i}{n\Delta}.$$

Полигон можно получить из готовой гистограммы следующим образом: середины верхних сторон прямоугольников, образующих гистограмму, нужно соединить отрезками прямых.

Изображая на одном рисунке несколько полигонов одновременно, мы можем сделать анализ процесса или явления в разных условиях.

Полигон, как и гистограмма, является статистическим аналогом функции плотности для непрерывных распределений.

Для непрерывного признака предпочтительнее использовать гистограмму, а для дискретного – полигон.

Кумулятой называется ломаная кривая, соединяющая точки $(x_1, h_1), (x_2, h_2), \dots, (x_r, h_r)$, где в качестве высоты h_i принимаются значения накопленных частот (n_i^* или w_i^*).

Существуют два варианта построения кумуляты.

1) По типу полигона. Кумулята представляет собой ломанную линию, которая соединяет вершины ординат, соответствующие накопленным частотам для данного значения признака.

2) По типу гистограммы. Кумулята представляет собой ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат интервалы, а высоты равны накопленным частотам.

Кумулята является статистическим аналогом функции распределения случайной величины.

► **Пример.** По результатам исследования на тему «Сколько Вы тратите на мобильную связь в месяц?» были получены следующие данные (руб.):

131	285	46	97	125	78	275	190	244
142	52	204	210	180	391	167	356	259
356	156	381	60	231	284	427	148	507
264	453	246	402	222	346	311	56	235

Исследование проводилось среди респондентов возрастной группы 25–30 лет.

Представьте данные в виде таблицы интервального ряда. Постройте гистограмму и полигон относительных частот, кумуляту.

Решение.

1. Всего наблюдений $n = 36$. Определим ширину интервала. Поскольку в исследовании речь идет о денежных суммах, причем, максимальное значение равно 507 рублей, то выберем интервалы одинаковой ширины – 100 рублей. То есть, $\Delta = 100$.

2. Найдем границы интервалов.

$a_0 = 0$, $a_1 = a_0 + \Delta = 0 + 100 = 100$, $a_2 = a_1 + \Delta = 100 + 100 = 200$ и т.д.

3. Найдем частоту n_i – число наблюдений, попавших в i -й интервал (a_{i-1}, a_i) , $i=1..7$. Например, 5-му интервалу (400–500) принадлежат 3 значения: 402, 427 и 453. Следовательно, $n_5 = 3$. Полученные значения занесем в таблицу 3 (с. 17).

4. Построим гистограмму. Каждый i -й интервал «достроим» до прямоугольника: основанием прямоугольника будет сам интервал (a_{i-1}, a_i) , а высота равна $\frac{n_i}{n\Delta}$. Например, над интервалом (100; 200) по-

строим прямоугольник, параллельный оси ОУ, высотой

$$h_3 = \frac{n_3}{n\Delta} = \frac{8}{36 \cdot 100} = 0.0022 \text{ (рис. 1, с. 17)}$$

Таблица 3

Номер интервала	Границы интервала (a_{i-1}, a_i)	Середина интервала, v_i	Частота n_i	Высота $h_i = \frac{n_i}{n\Delta}$	Накопленная частота n_i^*
1	0–100	50	6	0,00167	6
2	100–200	150	8	0,0022	14
3	200–300	250	12	0,00333	26
4	300–400	350	6	0,00167	32
5	400–500	450	3	0,00083	35
6	500–600	550	1	0,00028	36

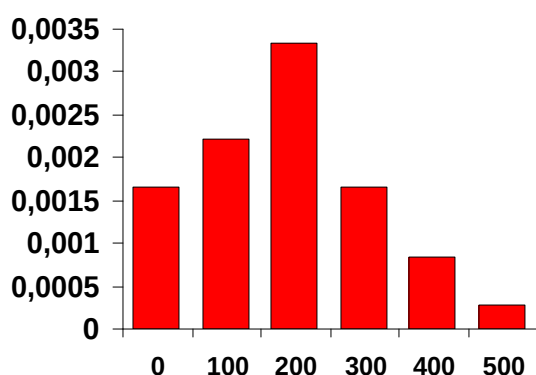


Рис. 1. Гистограмма

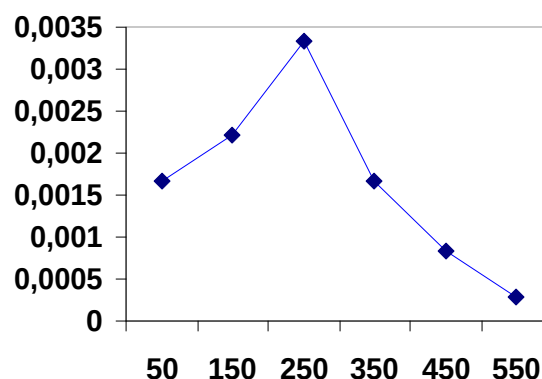


Рис. 2. Полигон

5. Построим полигон.

а) Найдем середины интервалов по формуле

$$v_i = \frac{a_{i-1} + a_i}{2}.$$

$$v_1 = \frac{a_0 + a_1}{2} = \frac{0 + 100}{2} = 50 \text{ и т.д.}$$

б) Для построения полигона соединим отрезками точки (v_i, h_i) , $i=1...6$: (50; 0,00167), (150; 0,0022), (250; 0,00333), (350; 0,00167), (450; 0,00083), (550; 0,00028) (рис. 2).

6. Построим кумуляту.

а) Найдем накопленные частоты n_i^* .

$n_1^* = 6$; $n_2^* = 6 + 8 = 14$; $n_3^* = 6 + 8 + 12 = 26$; $n_4^* = 6 + 8 + 12 + 6 = 32$ и т.д.

б) Для построения кумуляты каждый i -й интервал «достроим» до прямоугольника: основанием прямоугольника будет сам интервал (a_{i-1}, a_i) , а высота равна n_i^* . Так, над интервалом $(100; 200)$ построим прямоугольник, параллельный оси ОУ, высотой $n_i^* = 14$ (рис. 3).

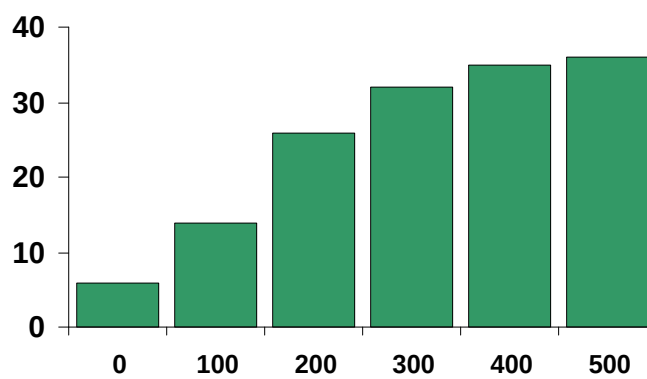


Рис. 3. Кумулята

Проанализируем полученные результаты. Самые большие расходы на мобильную связь в месяц составили 507 рублей, а самые низкие – 46 рублей. Чаще всего респонденты тратили на мобильную связь от 200 до 300 рублей в месяц. Это треть всех опрошенных (12 человек). В общем 72,2 % (26 человек) потратили на мобильную связь сумму, не превышающую 300 рублей. ■

Контрольные вопросы

1. Как строится гистограмма?
2. Каким образом следует выбирать интервалы группировки при построении гистограммы?
3. Сколько интервалов нужно выбрать?

4. Как связаны значения гистограммы и функции плотности?
5. Как строится полигон?
6. В каких случаях предпочтительнее строить гистограмму, а в каких – полигон?

Задачи к §2.1

1. Для изучения потенциального рынка мобильных телефонов задавался следующий вопрос: «Сколько лет Вашему мобильному телефону, т.е. телефону, которым Вы сейчас пользуетесь?» Были получены следующие результаты: 3, 6, 3, 1, 3, 4, 3, 5, 3, 2, 3, 1, 3, 3, 5, 4, 2, 4, 4, 2, 4, 4, 1, 4, 2, 4, 2, 4. Составить вариационный ряд, таблицу частот. Построить полигон относительных частот и кумуляту.

2. Владелец кафе для определения распределение размера заказов, сделанных в будний день в обеденное время (с 13.00 до 15.00), случайным образом выбрал чеки на подобные заказы. Результаты отражены в следующей таблице.

Размер заказа, рублей	0–50	50– 100	100– 150	150– 200	200– 250	250– 300	300– 350	350– 400
Число заказов, ед.	2	3	11	25	28	24	15	4

Построить гистограмму относительных частот.

3. Составьте таблицу частот для следующего распределения размеров 43 пар мужской обуви, проданных магазином за день: 39, 41, 40, 42, 41, 40, 42, 44, 40, 43, 42, 41, 43, 39, 42, 41, 42, 39, 41, 37, 43, 41, 38, 43, 42, 41, 40, 41, 38, 44, 40, 39, 41, 40, 42, 40, 41, 42, 40, 43, 38, 39, 44. Какой размер обуви является наиболее покупаемым, и какая доля покупателей имеет этот размер? Постройте а) полигон частот; б) кумуляту, взяв в качестве высоты n_i^* .

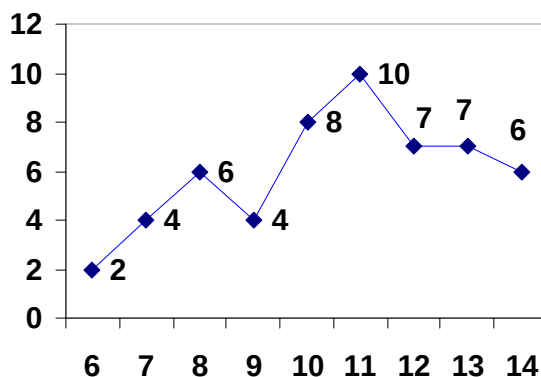
4. Для исследования успеваемости студентов была проведена контрольная работа по статистике (максимум = 18 баллов). Ее результаты:

11, 16, 18, 12, 9, 4, 12, 14, 8, 15, 16, 18, 15, 14, 11, 12, 14, 7, 10 баллов. Проведите группировку полученных оценок, используя для определения границ интервалов следующую таблицу.

Оценка	Баллы
Неудовлетворительно	0–9
Удовлетворительно	10–12
Хорошо	13–15
Отлично	16–18

Постройте а) гистограмму частот; б) полигон частот. Какая оценка встречается чаще всего? Определите долю студентов, получивших именно эту оценку.

5. По следующему полигону постройте кумуляту.



6. Динамика регистрации разводов в зависимости от прожитых лет совместной жизни (% , Татарстан)

Год	Количество прожитых лет совместной жизни			
	0–5	5–10	10–15	15 и выше
2011	41	24	11	24
2012	35	24,5	12	28,5
2013	36	26	12	26

1) Постройте гистограмму частот по данным а) 2011 г.; б) 2012 г.

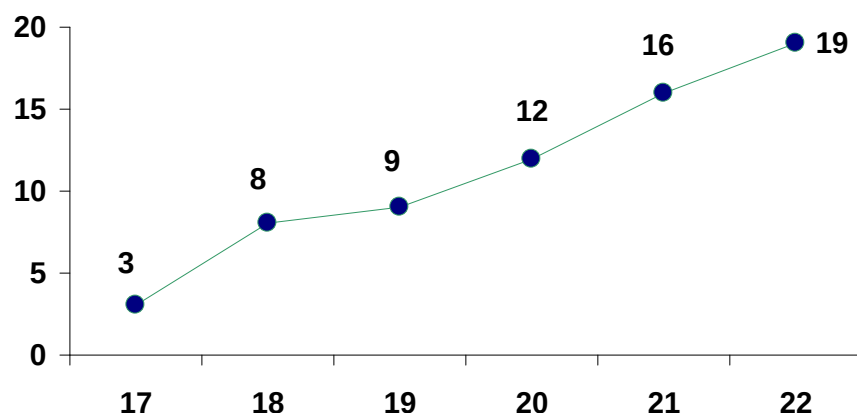
2) Изобразите на одной плоскости полигоны относительных частот по данным 2011 – 2013 гг.

3) Постройте кумуляту по данным 2013 г.

7. Постройте гистограмму относительных частот по первичным данным размера прибыли кредитных организаций области за год (млн д.е.).

3,5	4,8	2,5	3,9	8,7	9,4	10,5	4,9	12,8	6,7	8,9	4,6	2,8	3,4	5,7
8,9	5,4	2,3	5,6	3,7	4,2	5,4	8,9	8,8	5,7	11,2	10,8	9,4	12,5	7,1
10,	8,4	5,6	3,7	1,8	4,9	5,7	3,7	6,4	8,7	4,9	3,5	7,6	6,4	3,8

8. По следующему графику кумуляты восстановите исходные данные и постройте полигон.



9. В результате исследования затрат времени на обработку одной детали получены следующие результаты:

1.9 3.1 1.3 1.1 0.7 2.9 2.7 4.0 1.7 3.2 0.9 0.8 1.2 2.6 1.9 2.3
 4.1 1.3 0.4 4.5 0.9 1.4 1.6 2.2 3.1 1.5 1.1 2.3 4.3 0.7 4.2 1.5
 1.8 2.9 0.8 0.9 1.7 4.1 2.6 0.9 0.8 1.2 1.9 3.2 2.9 1.1 3.2 2.1

а) Построить гистограмму частот по выборке, предварительно проведя группировку. (Указание. В качестве длины интервала можно взять следующие значения: а) $\Delta = 0,3$; б) $\Delta = 0,5$; в) $\Delta = 1,2$).

б) Построить кумуляту, взяв в качестве высоты w_i^* .

10.

Распределение численности населения Российской Федерации по полу и возрастным группам на 1 января 2015 г.

(тыс. человек)

Возраст, лет	Все население		
	мужчины и женщины	мужчины	женщины
Все население	146267	67772	78495
в том числе в возрасте, лет:			
до 1 года	1930	993	937
1–4	7332	3765	3567
5–9	8004	4098	3906
10–14	7126	3651	3475
15–19	6829	3497	3332
20–24	9293	4744	4549
25–29	12620	6379	6241
30–34	12092	6043	6049
35–39	10884	5327	5557
40–44	10122	4909	5213
45–49	9140	4352	4788
50–54	10957	5057	5900
55–59	10873	4774	6099
60–64	9260	3814	5446
65–69	6428	2479	3949
70 и более	13377	3890	9487

1) Постройте гистограмму относительных частот по данным всего населения (и мужчин, и женщин).

2) Изобразите на одной плоскости полигоны относительных частот по данным мужчин и женщин.

2.2. Числовые характеристики выборки

Ключевые слова: среднее арифметическое, мода, медиана, размах вариации, дисперсия, стандартное отклонение, коэффициент вариации.

Вариационные ряды и графики эмпирических распределений дают наглядное представление о том, как изменяется признак в выборочной совокупности. Но они не являются достаточными для полной характеристики выборки. Числовые характеристики выборки дают количественное представление об эмпирических данных и позволяют их сравнивать между собой. Наибольшее практическое применение имеют характеристики положения и характеристики вариации эмпирических распределений.

2.2.1. Характеристики положения

В этом разделе мы рассмотрим характеристики положения, определяющие положение центра эмпирического распределения.

1. *Среднее $\bar{x}$ (среднее арифметическое).*

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad (1)$$

где n – объем выборки; x_i – варианты выборки.

Для сгруппированных данных формулы среднего приведены в таблице 1 (с. 27).

Среднее показывает, какие значения признака наиболее характерны для данной совокупности в целом.

2. *Медианой ($\hat{Me}$)* называется значение признака, которое приходится на центральный член вариационного ряда.

Таким образом, одна половина вариационного ряда меньше (или равна) медианы, а вторая – больше (или равна).

$$\hat{Me} = \begin{cases} x_{k+1}, & n = 2k + 1 \\ \frac{x_k + x_{k+1}}{2}, & n = 2k \end{cases}, \quad (2)$$

где n – объем выборки

Для интервального ряда сначала определяется медианный интервал, потом по формуле вычисляется медиана.

а) *Медианным* будет тот интервал, в котором сумма частот n_i впервые окажется больше $\frac{\sum n}{2}$:

$$\begin{cases} n_1 + n_2 + \dots + n_{k-1} < \frac{\sum n}{2} \\ n_1 + n_2 + \dots + n_k > \frac{\sum n}{2} \end{cases}, \text{ т.е. } k\text{-й интервал будет медианным.}$$

б) Медиана для интервального ряда:

$$\hat{Me} = x_{Me} + \Delta \frac{\frac{\sum n}{2} - S_{Me-1}}{n_{Me}}, \quad (3)$$

где x_{Me} – нижняя граница медианного интервала,

Δ – величина медианного интервала,

S_{Me-1} – сумма накопленных частот, предшествующих медианному интервалу,

n_{Me} – частота медианного интервала.

Медиану можно найти по графику кумуляты (рис. 1, с. 25). Для этого следует провести горизонтальную линию $y = 0,5$ (при $h_i = w_i^*$) или $y = \frac{n}{2}$ (при $h_i = n_i^*$) до пересечения с графиком кумуляты.

Медиана применяется при статистическом контроле качества продукции и технологического процесса на предприятии; при изучении распределения семей по величине дохода и т. п.

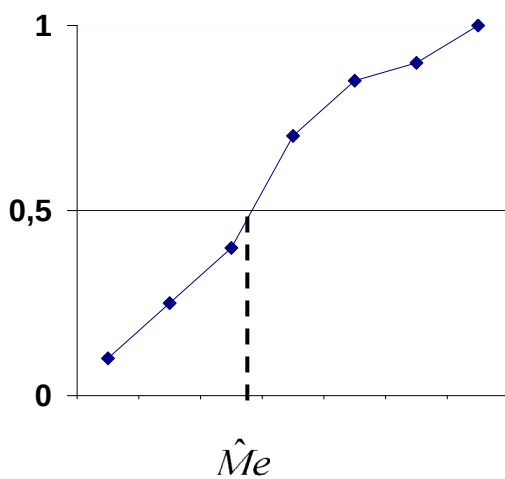


Рис. 1. Кумулята

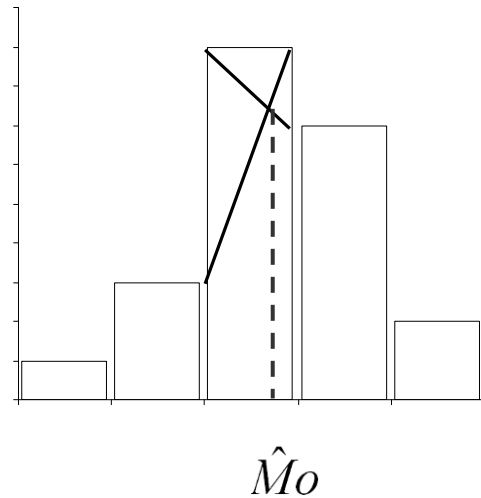


Рис. 2. Гистограмма

3. *Модой* ($\hat{M}_o$) называется наиболее часто встречающееся в выборке значение признака.

Для дискретного ряда мода определяется визуально по максимальной частоте или частости.

Для интервального ряда сначала определяется модальный интервал, потом по формуле вычисляется мода.

Модальным называется интервал, имеющий наибольшую частоту (частость).

$$\hat{M}_o = x_{M_o} + \Delta \frac{n_{M_o} - n_{M_o-1}}{(n_{M_o} - n_{M_o-1}) + (n_{M_o} - n_{M_o+1})}, \quad (4)$$

где x_{M_o} – нижняя граница модального интервала,

Δ – величина модального интервала,

n_{M_o} – частота модального интервала,

n_{M_o-1} – частота интервала, предшествующего модальному,

n_{M_o+1} – частота интервала, следующего за модальным.

Моду также можно найти по графику гистограммы, как показано на рис. 2.

Совокупность может иметь одну моду (унимодальное распределение) или несколько мод (мультимодальное распределение).

Мода широко используется при изучении покупательского спроса, ходовых размеров одежды и обуви, регистрации цен и т.д.

Мода и медиана, в отличие от среднего, мало чувствительны к колебаниям крайних вариантов. Поэтому, если эмпирическое распределение сильно асимметрично, то предпочтительнее использовать моду и медиану. Вообще, среднее, мода и медиана совпадают только в случае симметричного и унимодального (с одним максимумом) распределения.

В статистических исследованиях применяются различные виды средних величин. Выбор средней должен осуществляться с учетом содержания исследуемого показателя и формы представления данных.

2.2.2. Характеристики вариации

Средние величины не дают полной информации о варьирующем признаке. При одинаковых средних признаки могут отличаться по величине и разбросу значений. Поэтому наряду со средними значениями вычисляют и характеристики вариации (рассеяния) выборки.

1. *Размах вариации* R показывает разность между наибольшими и наименьшими значениями выборки:

$$R = x_{\max} - x_{\min} . \quad (5)$$

2. Дисперсия и стандартное отклонение являются важнейшими характеристиками вариации.

Дисперсией S^2 называется средний квадрат отклонения значений вариант от их средней арифметической:

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 . \quad (6)$$

Дисперсия после преобразования принимает вид:

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2. \quad (7)$$

Если массив данных был подвергнут группировке, то выборочная дисперсия вычисляется по формулам таблицы 1.

Таблица 1

Данные	Среднее	Дисперсия
$x_1, x_2, \dots, x_n$	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$	$S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2$
Дискретный вариационный ряд	$\bar{x} = \frac{1}{n} \sum_{j=1}^r x_j n_j$	$S^2 = \frac{1}{n} \sum_{j=1}^r (x_j - \bar{x})^2 n_j = \frac{1}{n} \sum_{j=1}^r x_j^2 n_j - (\bar{x})^2$
Интервальный вариационный ряд	$\bar{x} = \frac{1}{n} \sum_{j=1}^r v_j n_j$	$S^2 = \frac{1}{n} \sum_{j=1}^r (v_j - \bar{x})^2 n_j = \frac{1}{n} \sum_{j=1}^r v_j^2 n_j - (\bar{x})^2$

$n = \sum_{j=1}^r n_j$ – объем выборки; r – число интервалов группировки;

x_j – варианты дискретного ряда; v_j – середина j -го интервала интервального вариационного ряда;

n_j – частота варианты x_j или j -го интервала.

Стандартное отклонение (или среднее квадратическое отклонение) определяется как корень квадратный из дисперсии:

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}. \quad (8)$$

Стандартное отклонение имеет ту же размерность, что и сам признак. Поэтому на практике обычно используется стандартное отклонение, а не дисперсия.

3. Коэффициент вариации

Выборочное стандартное отклонение выражается в тех же единицах измерения, что и характеризующий им признак. В практике же бывают ситуации, когда сравнивают изменчивость признаков, выраженных разными единицами. В таких случаях используют не абсолютные, а относительные показатели вариаций. Одним из таких показателей является *коэффициент вариации*:

$$C_v = \frac{S}{\bar{x}} 100\% . \quad (9)$$

Если $C_v \leq 10 \%$, то варьирование считается слабым, при $11\% \leq C_v \leq 25\%$ – средним, и значительным при $C_v > 25\%$. Если коэффициент вариации не превышает 10 %, то выборку можно считать однородной, т.е. полученной из одной генеральной совокупности.

Следует заметить, что к понятиям среднее, дисперсия, стандартное отклонение часто добавляются слова «эмпирическое» или «выборочное». Тем самым подчеркивается, что указанные характеристики вычислены по экспериментальным, опытным данным.

2.2.3. Упрощенные методы расчета выборочной средней и выборочной дисперсии

В тех случаях, когда экспериментальные данные x_i представлены большими числами, вычисление среднего и, особенно, дисперсии усложняется громоздкими операциями. Чтобы избежать этого, перейдем к условным вариантам $u_i = x_i - A$. В качестве A выберем число, близкое к выборочному среднему, или любое другое число по нашему усмотрению. Возможна другая ситуация, когда исходные данные являются десятичными дробями с k десятичными знаками после запятой. В этом случае удобно перейти к условным вариантам вида: $u_i = \frac{x_i}{C}$, где $C = 10^{-k}$. Подобная замена используется и в случае, когда

исходные данные имеют следующий вид: 1000, 3000, 2000, 4000. Приняв $C = 10^3$, получим преобразованные данные: 1, 3, 2, 4.

Иногда оказывается полезным совместное использование двух рассмотренных выше примеров, например, преобразование вида:

$$u_i = \frac{x_i - A}{C}.$$

Как изменятся среднее и дисперсия при таких преобразованиях? Формулы представлены в таблице 2.

Таблица 2

Замена	Среднее, $\bar{x}$	Дисперсия, S_x^2
$u_i = x_i - A$	$\bar{x} = \bar{u} + A$	$S_x^2 = S_u^2$
$u_i = \frac{x_i}{C}$	$\bar{x} = C \bar{u}$	$S_x^2 = C^2 S_u^2$
$u_i = \frac{x_i - A}{C}$	$\bar{x} = C \bar{u} + A$	$S_x^2 = C^2 S_u^2$

Значения $\bar{u}$ вычисляются по формуле (1): $\bar{u} = \frac{1}{n} \sum_{i=1}^n u_i$,

а S_u^2 – по формуле (4): $S_u^2 = \frac{1}{n} \sum_{i=1}^n (u_i - \bar{u})^2 = \frac{1}{n} \sum_{i=1}^n u_i^2 - (\bar{u})^2$.

► **Пример 1.** Трудовой стаж рабочих бригады равен соответственно: 1, 3, 6, 8, 3, 10, 5 (лет). Найти средний трудовой стаж, медиану, дисперсию и стандартное отклонение.

Решение. Число сотрудников равно $n = 7$.

1) Средний трудовой стаж находим по формуле среднего арифметического (1):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1+3+6+8+3+10+5}{7} = \frac{36}{7} = 5,14 \text{ лет}.$$

2) Для определения медианы сначала построим вариационный ряд:

1, 3, 3, 5, 6, 8, 10.

Найдем медиану двумя способами.

Способ 1 – визуальный.

Медиану, как середину вариационного ряда, можно определить зрительно – это значение «5».

$\hat{Me}$

1, 3, 3, **5**, 6, 8, 10.

Способ 2 – по формуле (2).

$n = 7$ – число нечетное, т.е. $n = 2k + 1 = 7$. Следовательно, $k = 3$.

$\hat{Me} = x_{k+1} = x_{3+1} = x_4 = 5$ лет – это трудовой стаж рабочего с условным номером 4. Из оставшихся у половины рабочих (№№ 1–3) трудовой стаж меньше медианного, а у другой половины (№№ 5–7) – больше.

3) Дисперсию можно найти двумя способами.

Способ 1. Используем определение дисперсии (6):

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{7} ((1 - 5,14)^2 + (3 - 5,14)^2 + (6 - 5,14)^2 + (8 - 5,14)^2 + (3 - 5,14)^2 + (10 - 5,14)^2 + (5 - 5,14)^2) = 8,4$$

Стандартное отклонение равно $S = \sqrt{S^2} = \sqrt{8,4} = 2,9$.

Способ 2. Используем упрощенную формулу выборочной дисперсии (7):

$$S^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2 = \frac{1}{7} (1^2 + 3^2 + 6^2 + 8^2 + 3^2 + 10^2 + 5^2) - (5,14)^2 = 8,4.$$

Способ 2 требует меньше вычислений. Но, если мы оперируем большими числами, то предпочтительнее будет использовать способ 1 или перейти к условным вариантам (§2.3). ■

► **Пример 2.** Руководство компании, обеспокоенное постоянными опозданиями сотрудников с обеденного перерыва, решило провести исследование этой проблемы. Наблюдения проводились в среду, т.к. этот день недели считается одним из наиболее спокойных и продуктивных дней недели. Опоздания на 1–2 минуты было решено не фиксировать. В результате были получены следующие данные.

Время опоздания X , мин.	3	4	5	6	9	10	15	25
Количество людей n_i	6	6	20	13	5	8	1	1

Вычислите среднее времени опоздания, модальное значение времени опоздания, размах вариации, дисперсию и стандартное отклонение, коэффициент вариации.

Решение. Объем выборки равен

$$n = \sum n_i = 6 + 6 + 20 + 13 + 5 + 8 + 1 + 1 = 60.$$

1) Найдем среднее времени опоздания. Данные представлены в виде дискретного вариационного ряда, поэтому будем использовать формулу среднего для дискретного вариационного ряда (таблица 1):

$$\bar{x} = \frac{1}{n} \sum_{j=1}^7 x_j n_j = \frac{1}{60} (3 \cdot 6 + 4 \cdot 6 + 5 \cdot 20 + 6 \cdot 13 + 9 \cdot 5 + 10 \cdot 8 + 15 \cdot 1 + 25 \cdot 1) = 6,42 \text{ мин.}$$

2) Найдем моду: $\max n_i = n_3 = 20$. Тогда $\hat{M}_0 = x_3 = 5 \text{ мин}$. Следовательно, чаще всего сотрудники опаздывали на 5 мин. Это 20 человек из 60, т.е. треть сотрудников опоздали ровно на 5 мин.

3) Наименьшее время опоздания было 3 мин, а наибольшее – 25 мин.

По формуле (5) находим размах вариации времени опоздания

$R = x_{\max} - x_{\min} = 25 - 3 = 22 \text{ мин}$. Найденное значение размах вариации не совсем объективно отражает вариацию времени опоздания сотрудников с обеда. Для большей части сотрудников (58 человек из 60) время опоздания составляет максимум 10 мин. И только 2 человека превысили это время: один опоздал на 15 мин, а другой – на 25 мин.

Здесь проявился недостаток характеристики «размах вариации», которая чувствительна к значениям крайних вариантов.

4) Дисперсию будем вычислять по упрощенной формуле дисперсии для дискретного вариационного ряда (таблица 1):

$$S^2 = \frac{1}{n} \sum x_i^2 n_i - (\bar{x})^2 = \frac{1}{10} (3^2 \cdot 6 + 4^2 \cdot 6 + 5^2 \cdot 20 + 6^2 \cdot 13 + 9^2 \cdot 5 + 10^2 \cdot 8 + 15^2 \cdot 1 + 25^2 \cdot 1) - 6,42^2 = 11,7 \text{ мин}^2$$

5) Стандартное отклонение равно $S = \sqrt{S^2} = \sqrt{11,7} = 3,42$ мин, т.е. конкретные значения времени опоздания в среднем отклоняются от их среднего значения на 3,42 мин.

6) Вычислим коэффициент вариации по формуле (9):

$$C_v = \frac{S}{\bar{x}} 100\% = \frac{3,42}{6,42} 100\% = 53,32\%.$$

Варьирование признака является значительным, что объясняется неоднородностью выборки.

Анализируя полученные результаты, мы можем сделать следующие выводы. Среднее значение времени опоздания составило 6,42 мин., а наиболее частое время опоздания – 5 мин. Разница во времени опоздания была 22 мин. В целом вариацию можно охарактеризовать как значительную. ■

▶ Пример 3.

Сколько времени Вы тратите на просмотр телепередач
в будний день?

Время X, час	0–0,5	0,5–1	1–3	3–6	6–9	9–15	Более 15
Число людей n_i	4	13	38	15	4	2	1

Вопрос задавался тем респондентам, которые смотрят телевизор.

Найти среднее и стандартное отклонение.

Решение. Всего в опросе участвовало

$$n = \sum n_i = 4+13+38+15+4+2+1 = 77 \text{ человек.}$$

Преобразуем данный интервальный вариационный ряд в безынтервальный. Для этого в качестве вариант возьмем середины интервалов. Например, для 3-го интервала середина будет $v_3 = \frac{1+3}{2} = 2$. У последнего интервала нет верхней границы, поэтому мы её выберем самостоятельно. Мы возьмем возможный максимум – 24 часа. В результате таких преобразований мы получаем следующую таблицу.

Время V , час	0,25	0,75	2	4,5	7,5	12	19,5
Число людей n_i	4	13	38	15	4	2	1

1) Вычислим среднее.

$$\bar{x} = \frac{1}{n} \sum_{j=1}^n v_j n_j = \frac{1}{77} (0,25 \cdot 4 + 0,75 \cdot 13 + 2 \cdot 38 + 4,5 \cdot 15 + 7,5 \cdot 4 + 12 \cdot 2 + 19,5 \cdot 1) = 2,96 \text{ час.}$$

2) Вычислим стандартное отклонение.

$$S^2 = \frac{1}{n} \sum_{j=1}^n v_j^2 n_j - (\bar{v})^2 = \frac{1}{77} (0,25^2 \cdot 4 + 0,75^2 \cdot 13 + 2^2 \cdot 38 + 4,5^2 \cdot 15 + 7,5^2 \cdot 4 + 12^2 \cdot 2 + 19,5^2 \cdot 1) - (2,96)^2 = 8,87$$

Стандартное отклонение равно $S = \sqrt{S^2} = \sqrt{8,87} = 2,98 \text{ час.}$

Итак, мы нашли, что среднее время, проведенное у телевизора, составляет почти 3 часа (2,96 ч), а стандартное отклонение – 2,98 ч.

Замечание. Предположим, что мы в качестве верхней границы последнего интервала взяли не 24 часа, а 17 часов. Это значение мы определили, исходя из того, что по медицинским нормам на сон взрослому человеку требуется 7–8 часов. Тогда мы бы получили $\bar{X} = 2,91$; $S^2 = 7,52$; $S = 2,74$, которые не намного отличаются от предыдущих расчетов. Это связано с тем, что в последний интервал попадает только 1 из 77 наблюдений, поэтому этот интервал не вносит значимого вклада в расчеты характеристик. ■

Контрольные вопросы

1. Дайте определение средней.
2. Какие виды средних величин применяются в математической статистике?
3. Почему среднее арифметическое интервального ряда является приближенной средней, от чего зависит степень ее приближения?
4. Дайте определение медианы.
5. Дайте определение дисперсии.
6. Что называется стандартным отклонением? По каким формулам оно вычисляется?

Задачи к §2.2

1. Продолжительность больничных листов, оформленных сотрудниками организации в период эпидемии гриппа, составила: 6, 3, 4, 7, 6, 12, 3 (дня). Найдите а) среднюю продолжительность больничных листов, б) медиану, в) моду. Определите размах вариации и выборочную дисперсию.
2. За неделю в населенном пункте было зафиксировано следующее количество ДТП (по дням недели).

Понедельник	Вторник	Среда	Четверг	Пятница	Суббота	Воскресенье
6	3	2	1	2	4	8

Найдите а) среднее число ДТП, б) медиану, в) моду, г) стандартное отклонение.

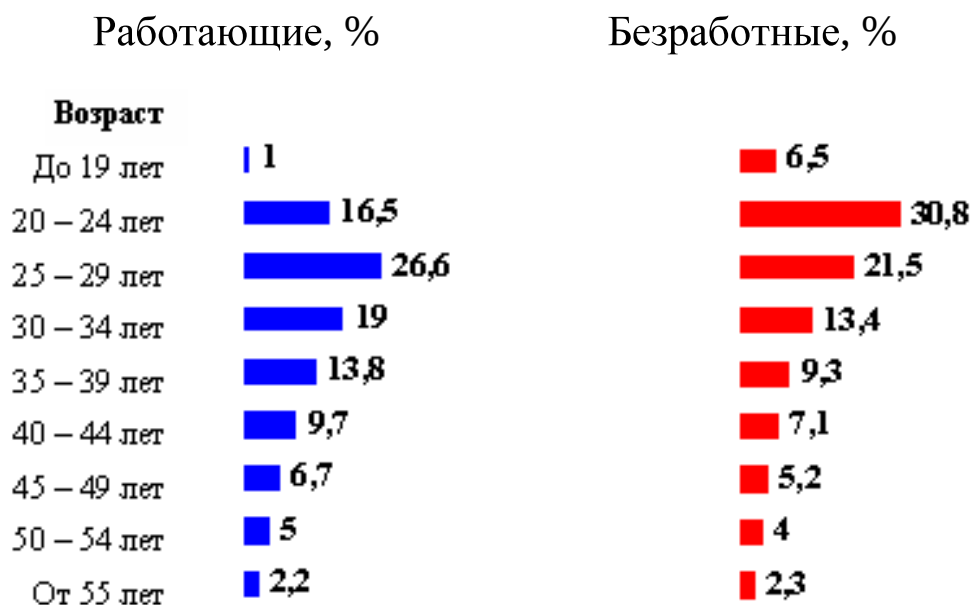
3. Выборочное исследование домохозяйств области дало следующие результаты.

	Состав домохозяйств, человек					
	1	2	3	4	5	6 и более
Число домохозяйств, ед.	32	32	29	15	7	5

Найдите модальный и медианный размеры домохозяйства.

4. Исследовательский центр портала Superjob.ru составил следующий портрет соискателя (октябрь 2013 г.).

Портрет соискателя



Найдите медианный и модальный возрасты а) работающего соискателя; б) безработного соискателя.

5. В ходе маркетингового исследования выяснялось, в частности, насколько новая упаковка является привлекательной для покупателей. Посетителям магазина было предложено оценить по 5-балльной шкале новый дизайн упаковки. Были получены следующие результаты.

Оценка, балл	0	1	2	3	4	5
Количество покупателей	1	2	4	13	24	12

Сколько человек участвовало в исследовании? Определите средний балл качества продукции, коэффициент вариации. Какая оценка является модальной?

6. Найдите стандартное отклонение, если для 116 наблюдений сумма квадратов равна 3729, среднее арифметическое – 5,25.

7. Найдите дисперсию, если коэффициент вариации равен 20,9%, а среднее арифметическое – 8,91.

8. В таблице представлены данные о возрасте молодых людей, состоящих в зарегистрированном браке.

	Городское и сельское население		Городское население		Сельское население	
	Женщины	Мужчины	Женщины	Мужчины	Женщины	Мужчины
15–19	4,7	0,9	4,2	0,9	6,8	1,0
20–24	34,3	18,6	32,6	18,3	40,2	19,6
25–29	56,1	48,6	54,5	48,5	61,4	48,7

(Росстат, «Молодежь России 2010»)

1. Найдите средний возраст молодой женщины, состоящей в зарегистрированном браке а) среди городского и сельского населения; б) среди городского населения; в) среди сельского населения.
2. Найдите средний возраст молодого мужчины, состоящего в зарегистрированном браке среди городского и сельского населения.

9. В рамках выборочного обследования об организации досуга, проведенным Росстатом среди молодежи, были получены следующие данные.

Время, затраченное на занятия физкультурой и спортом
(в % от числа респондентов; по данным выборочного обследования 2008 г.)

Время, затраченное на занятия физкультурой и спортом	Возрастные группы, лет		
	15–19	20–24	25–29
До 1 часа в неделю	4,8	13,6	19,5
От 1 до 3 часов	26,7	40,8	43,7
От 3 до 6 часов	42,9	24,8	23,0
От 6 до 9 часов	10,5	12,0	5,7
От 9 до 12 часов	7,6	4,8	6,9
От 12 до 15 часов	4,8	0,8	—
Свыше 15 часов	2,7	3,2	1,2

(Росстат, «Молодежь России 2010»)

Определите среднее время, затраченное на занятия физкультурой и спортом в каждой возрастной группе. Вычислите стандартные отклонения в каждой возрастной группе. (Указание. В качестве верхней границы в интервале «Свыше 15 часов» предлагается взять 25 часов.)

10. В летний уборочно-заготовительный период семья Ивановых для консервирования приобрела сахар по следующим ценам: 10 кг по 25 д.е., 5 кг по 26,30 д.е., 7 кг по 30 д.е. и 20 кг им удалось купить по 21 д.е. По какой в среднем цене обошелся сахарный песок семье Ивановых? Найдите размах вариации цен.

11. К 10-летнему юбилею компании было подготовлено статистическое исследование о работе компании, ее успехах, потенциале и кадрах. В том числе и о стаже работников главного офиса и его филиала. По данным, представленным в виде полигона (рис. 1), рассчитайте средний стаж а) работников главного офиса; б) работников филиала; в) всех работников – офиса и филиала – двумя способами.

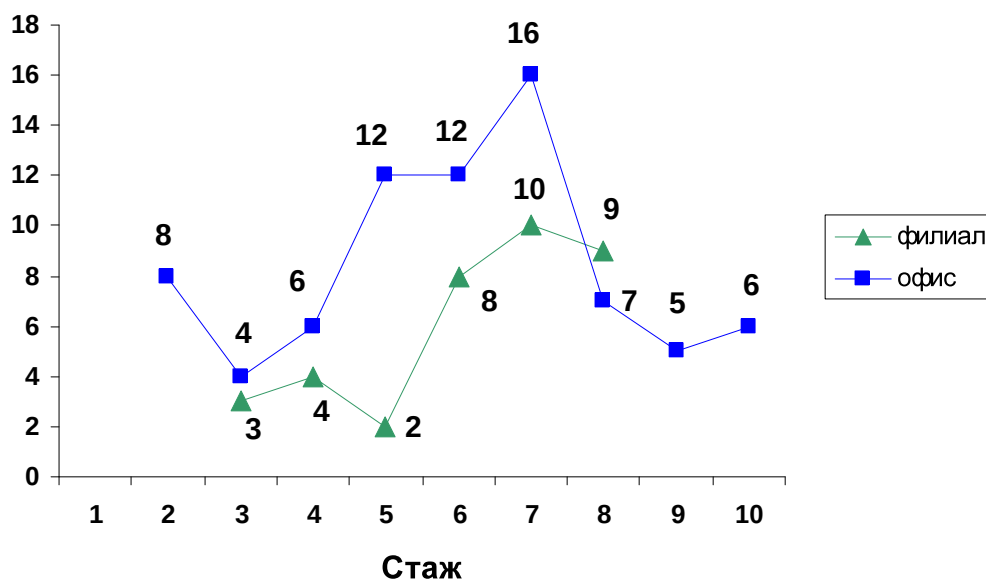


Рис. 1

12. По данным **Примера** (§2.1):

1) вычислить выборочное среднее и выборочное стандартное отклонение, используя группировку выборки;

2) вычислить среднее и стандартное отклонение не сгруппированной выборки, используя заданные значения вычисленных сумм:

$$\sum x_i = 8517, \quad \sum x_i^2 = 2534691.$$

Сравните полученные характеристики.

13. Выборочное исследование расхода электроэнергии по однокомнатным квартирам (2 проживающих) дало следующие результаты.

№	Месяц				
	ноябрь	декабрь	январь	февраль	март
1	538	586	638	678	715
2	376	428	492	544	592
3	854	921	999	1067	1131
4	400	435	473	508	549
5	63	125	200	271	326
6	140	185	234	275	319
7	280	334	386	437	487
8	438	487	552	587	628

Данные электросчетчика приведены на конец месяца.

Стоимость 1 кВт/ч равна 3,2 д.е.

1) Найти среднее потребление электроэнергии в форме выборочной средней а) за январь; б) за февраль; в) за 1-й квартал; г) за зимние месяцы.

2) Найти стандартное отклонение по данным потребления за февраль.

3) Найти среднемесячную оплату за электроэнергию а) за январь; б) за март.

14. Найдите коэффициент вариации, если при 64 наблюдениях $\sum x_i^2 = 6253$, $\sum x_i = 553$.

15. В компании были введены следующие прогрессивные штрафы за опоздания на работу.

Время опоздания, мин.	Штраф, д.е.
5–10	50
11–20	100
21–30	200
Более 30	Выговор

За опоздание менее 5 минут было решено пока не штрафовать. По результатам месяца была получена следующая статистика.

7 1 5 12 4 15 25 9 2 18 26 45 2 25 5
 14 8 18 6 22 7 4 13 6 25 16 3 32 51 9

- Найти а) среднее время опозданий;
 б) среднее время опозданий в каждой группе (по градации штрафов);
 в) общую сумму штрафов по каждой группе;
 г) общую сумму штрафов;
 д) средний размер штрафа.

16. Время на сборку прибора (в мин.):

56 72 66 58 63 68 51 70 68 62 68 69 72 73 75 77 76 74 77 91
 82 84 86 85 78 74 66 86 58 79 82 69 68 77 72 59 71 64 84 64

$$\Delta = 3, \quad \sum x_i = 2874, \quad \sum x_i^2 = 209826.$$

1) Вычислить среднее и дисперсию, предварительно проведя группировку выборки с заданной длиной интервала Δ ;

2) Вычислить среднее и дисперсию не сгруппированной выборки, используя заданные значения вычисленных сумм.

Сравнить результаты вычислений.

3. ВЫБОРОЧНЫЙ МЕТОД

Ключевые слова: сплошное наблюдение, несплошное наблюдение, выборка, статистические оценки, точечные оценки параметров распределения, интервальные оценки параметров распределения.

В практике статистических наблюдений существует 2 метода наблюдений: сплошное и несплошное. При сплошном наблюдении изучаются все объекты (перепись населения, референдум), а при несплошном – только часть объектов.

Выборочным называется такое несплошное наблюдение, при котором признаки регистрируются у отдельных единиц изучаемой статистической совокупности, отобранных с использованием специальных методов, а полученные в процессе обследования результаты с определенным уровнем вероятности распространяются на всю исходную совокупность.

Выборочный метод позволяет сократить время и средства на исследования, сводит к минимуму порчу и уничтожение исследуемых объектов, значительно снижает ошибки регистрации и в итоге дает достаточно большую точность результатов.

Недостатком выборочного метода являются ошибки исследования, которые называют ошибками репрезентативности.

3.1. Статистические оценки параметров распределения

Одна из основных задач математической статистики заключается в нахождении распределения случайной величины X (с.в.) по данным выборки. Очень часто вид распределения с.в. X известен с точностью до неизвестных параметров. В этом случае нам остается найти приближенные значения этих параметров. Допустим, $F(x; \theta)$ –

функция распределения с.в. X , содержащая неизвестный параметр θ , и пусть $x_1, x_2, \dots, x_n$, – выборка. Требуется определить, используя эти наблюдения, число, которое можно было бы взять в качестве значения θ , или интервал, о котором можно было бы утверждать, что он с вероятностью, близкой к 1, покрывает это значение. В первом случае будем говорить о точечной оценке, во втором – о доверительном интервале (рис.1).



Рис. 1

Задача оценки параметров состоит в получении наилучших в определенном смысле оценок параметров распределения на основе выборочных данных.

3.1.1. Точечные оценки

Ключевые слова: точечная оценка, статистика, состоятельность, несмещенность, эффективность.

Для нахождения оценки неизвестного параметра θ мы должны построить некоторую статистику $T(x_1, x_2, \dots, x_n)$, которая является приближением к θ . Такая статистика T называется *оценкой* и обозначается, обычно, $\hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n)$. В дальнейшем термин «статистика» будет часто использоваться, поэтому уточним его: *статистикой* будем называть любую функцию от наблюдений $T = T(x_1, x_2, \dots, x_n)$. Пусть

$x_1, x_2, \dots, x_n$, – реализация случайной выборки из распределения со средним значением μ и дисперсией σ^2 . Припишем каждому значению вес $1/n$. Тогда пара $(x_k, \frac{1}{n})$, $k = 1..n$, образует выборочное распределение. Все характеристики, подсчитанные по выборочному распределению, называются выборочными характеристиками. В частности, в качестве оценки для истинного среднего μ возьмем *выборочное среднее* $\bar{x}$ (§2.2.1). Но для оценки среднего μ можно также использовать моду и медиану. Какая же из них наилучшая? И в каком смысле это оценка наилучшая?

Обычно выделяют три основных свойства оценок, по которым мы определяем их близость к истинному значению.

1. *Состоятельность.*

Точечная оценка $\hat{\theta}$ называется *состоятельной*, если при неограниченном увеличении объема выборки ($n \rightarrow \infty$) она стремится к истинному значению параметра θ .

Другими словами, состоятельная оценка с ростом объема выборки приближается к истинному значению оцениваемого параметра.

Выборочное среднее $\bar{x}$ является состоятельной оценкой истинного среднего μ , а выборочная дисперсия S^2 – состоятельная оценка дисперсии σ^2 . Так же выборочное стандартное отклонение $S = \sqrt{S^2}$ есть состоятельная оценка для истинного стандартного отклонения σ .

2. *Несмещенность.*

Статистика $\hat{\theta}$ называется *несмещенной* оценкой параметра θ , если ее математическое ожидание равно самому оцениваемому параметру, т.е. $E\hat{\theta} = \theta$.

В случае несмещенной оценки отклонение оценки от истинного значения носит несистематический, случайный характер. Т.е. несмещенная оценка, изменяясь от эксперимента к эксперименту, принима-

ет значения, которые в среднем совпадают с истинным значением оцениваемого параметра.

Из приведенных выше оценок только выборочное среднее имеет свойство несмещенности. Для дисперсии σ^2 несмещенной оценкой является исправленная (несмещенная) выборочная дисперсия, вычисляемая по формуле:

$$S_{\text{несм}}^2 = \frac{n}{n-1} S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

где S^2 вычисляется по формуле (2), §2.2.2.

3. Эффективность.

Несмещенная оценка является *эффективной*, если она имеет наименьшую дисперсию по сравнению с другими несмещенными оценками того же параметра.

Выборочное среднее $\bar{x}$ и выборочная дисперсия S^2 являются эффективными оценками соответствующих параметров.

Контрольные вопросы

1. В чем состоит задача оценки параметров?
2. Какие требования предъявляются к оценкам параметров?
3. Дайте определение состоятельности оценки и поясните смысл этого определения.
4. Можно ли сказать, что состоятельная оценка лучше не состоятельной оценки?
5. Определите разницу в понятиях: «статистика» (оценка) и «параметр» распределения.
6. Дайте определение несмещенности оценки и поясните смысл этого определения.
7. Дайте определение эффективности оценки и поясните смысл этого определения.

Задачи к §3.1.1

1. Докажите, что а) выборочное среднее $\bar{x}$ является несмещенной оценкой истинного среднего μ ; б) относительная частота является несмещенной оценкой вероятности.
2. Докажите, что выборочная дисперсия S^2 является смещенной оценкой дисперсии σ^2 .
3. Предположим, что при выборочном обследовании ($n = 100$) оценка дисперсии σ^2 составила 1,28. При расчете была использована формула для смещенной оценки, т.е. в знаменателе было n вместо $n - 1$. Возможно ли найти значение несмещенной оценки дисперсии σ^2 ?
(Ответ: 1,3.)

3.1.2 Интервальные оценки

Ключевые слова: доверительный интервал, надежность, уровень значимости, предельная ошибка выборки.

В предыдущем разделе мы рассмотрели точечные оценки. Но если мы имеем дело с выборкой малого объема, то случайный характер величины θ может привести к значительному расхождению между θ и $\hat{\theta}$, т.е. к ошибкам. В этом случае параметр θ удобно оценивать некоторым интервалом.

Интервальной называют оценку, которая определяется двумя числами – концами интервалов.

Доверительным интервалом для параметра θ называют интервал $(\hat{\theta}_1, \hat{\theta}_2)$ со случайными концами, накрывающий истинное значение θ с вероятностью не меньшей $1 - \alpha$, т.е.

$$P(|\theta - \hat{\theta}| < \Delta) \geq 1 - \alpha.$$

Число $\gamma = 1 - \alpha$ называется *доверительным уровнем (надежностью, доверительная вероятность)*.

Число α называется *уровнем значимости*.

Интерпретация доверительного уровня: в среднем в $(1-\alpha)100\%$ случаев фиксированное, но неизвестное значение θ будет накрыто доверительным интервалом. Обычно задают надежность, равную 0,95; 0,99; 0,999 (т.е. $\alpha = 0,05; 0,01; 0,001$) и риск ошибиться в первом случае будет 1 раз на 20 испытаний, во втором – один раз на 100 испытаний и в третьем – один раз на 1000 испытаний.

Наибольшее отклонение оценки $\hat{\theta}$ от оцениваемого параметра θ , которое возможно с заданным доверительным уровнем γ , называется предельной ошибкой выборки Δ .

Очень часто доверительные интервалы параметра θ представлены в виде $(\hat{\theta} - \Delta; \hat{\theta} + \Delta)$.

Предельная ошибка выборки характеризует точность оценки: чем больше значение Δ , тем шире интервал и, следовательно, тем меньше точность.

1. Доверительный интервал для среднего значения нормального распределения при известной дисперсии:

$$\bar{x} - u_{\alpha} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + u_{\alpha} \frac{\sigma}{\sqrt{n}}, \quad (1)$$

где u_{α} – квантиль нормального распределения, который находится из уравнения $\Phi(u_{\alpha}) = 1 - \frac{\alpha}{2}$.

Для случая повторной выборки $\Delta = u_{\alpha} \frac{\sigma}{\sqrt{n}}$. В случае бесповторной выборки в эту формулу включается поправка. Но для простоты вычислений (с точки зрения учебного процесса) задачи данного раздела будут сформулированы только для случая повторной выборки.

2. Доверительный интервал для среднего значения нормального распределения при неизвестной дисперсии:

$$\bar{x} - t_{\alpha} \frac{s}{\sqrt{n-1}} \leq \mu \leq \bar{x} + t_{\alpha} \frac{s}{\sqrt{n-1}}, \quad (2)$$

где S^2 – выборочная дисперсия, t_{α} – квантиль распределения Стьюдента.

3. Доверительный интервал для медианы:

$$\hat{Me} - t_{\alpha} \sqrt{\frac{\pi}{2}} \frac{s}{\sqrt{n-1}} \leq Me \leq \hat{Me} + t_{\alpha} \sqrt{\frac{\pi}{2}} \frac{s}{\sqrt{n-1}} \quad (3)$$

где S^2 – выборочная дисперсия, t_{α} – квантиль распределения Стьюдента.

4. Доверительные интервалы для оценки дисперсии нормального распределения:

$$\frac{nS^2}{\chi_{\alpha}^2} \leq \sigma^2 \leq \frac{nS^2}{\chi_{1-\alpha}^2}, \quad (4)$$

где $\chi_{\alpha}^2, \chi_{1-\alpha}^2$ – квантили распределения χ^2 (хи-квадрат) с $\nu = n - 1$ степенями свободы.

5. Доверительный интервал для доли ($n > 100$):

$$w - u_{\alpha} \sqrt{\frac{w(1-w)}{n}} \leq p \leq w + u_{\alpha} \sqrt{\frac{w(1-w)}{n}}, \quad (5)$$

где w – относительная частота, u_{α} – квантиль нормального распределения, который находится из уравнения $\Phi(u_{\alpha}) = 1 - \frac{\alpha}{2}$.

Замечание. При достаточно больших объемах выборки ($n > 30$) распределение Стьюдента приближается к нормальному. Поэтому, если σ^2 неизвестна и $n > 30$, то доверительный интервал для среднего можно вычислить по формуле (1). При этом величина стандартного отклонения σ заменяется на $S_{несм}$.

► **Пример 1.** С целью исследования качества работы почтовой службы была произведена случайная выборка 144 квитанций на доставку посылок из города А в город Б. Среднее время доставки посылок составило 3,4 дня. Постройте с надежностью 0,95 доверительный интервал для среднего времени доставки посылки, если стандартное отклонение известно и равно 1,8 дня.

Решение. Т.к. стандартное отклонение известно, то доверительный интервал строится по формуле (1). По условию $n = 140$; $\sigma = 1,8$; $\bar{x} = 3,4$. Поскольку $\gamma = 0,95$, то $\alpha = 1 - \gamma = 1 - 0,95 = 0,05$. Найдем $u_{0,05}$ из уравнения $\Phi(u_{0,05}) = 1 - \frac{0,05}{2} = 0,975$. По таблице «Нормальное распределение» (приложение, таблица I) находим $u_{0,05} = 1,96$, т.е. $\Phi(1,96) = 0,975$.

Нормальное распределение
$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$$

x	Сотые доли				
	...	4	5	6	7
...		...	...	...	...
1,8		0,9671	0,9678	0,9686	0,9693
1,9		0,9738	0,9744	0,9750	0,9756
2,0		0,9793	0,9798	0,9803	0,9808

Рис. 1. Фрагмент таблицы «Нормальное распределение»

Подставляя эти данные в формулу, получаем:

$$3,4 - 1,96 \frac{1,8}{\sqrt{144}} \leq \mu \leq 3,4 + 1,96 \frac{1,8}{\sqrt{144}},$$

$$3,1 \leq \mu \leq 3,69.$$

То есть, с надежностью 0,95 искомый доверительный интервал для среднего времени доставки посылки равен (3,1; 3,69) дней.

Найдем точность оценки: $\Delta = u_{\alpha} \frac{\sigma}{\sqrt{n}} = 1,96 \frac{1,8}{\sqrt{144}} = 0,294$. ■

► **Пример 2.** Объем заказов, полученных фирмой за последний месяц составил (млн руб.): 7,8; 6,5; 8,2; 9,4; 5,8; 6,8; 8,1; 7,5; 7,5; 8,9; 9,1; 9,6. Определите границы 95 %-х интервалов для среднего, медианы и дисперсии объема заказов.

Решение. По формулам §2.2 находим выборочное среднее, медиану, выборочную дисперсию и выборочное стандартное отклонение: $\bar{x} = 7,93$, $\hat{Me} = 7,95$, $S^2 = 1,3$, $S = 1,14$.

Доверительная вероятность равна $\gamma = 0,95$. Тогда уровень значимости $\alpha = 1 - \gamma = 1 - 0,95 = 0,05$.

а) Найдем доверительный интервал для среднего объема заказов.

По таблице «Критические точки распределения Стьюдента» (таблица III приложения) при заданном уровне значимости $\alpha = 0,05$, помещенному в верхней строке таблицы (для двусторонней критической области), и числе степеней свободы $\nu = n - 1 = 12 - 1 = 11$ находим $t_{\alpha} = 2,2$ (рис. 2).

Критические точки распределения Стьюдента

Число степеней свободы ν	Уровень значимости α (двусторонняя критическая область)					
	0, 10	0,05	0,02	0,01	0,002	0,001
...	...	...	...	...	...	...
10	1,81	2,23	2,76	3,17	4,14	4,59
11	1,80	2,2	2,72	3,11	4,03	4,44
12	1,78	2,18	2,68	3,05	3,93	4,32

Рис. 2. Фрагмент таблицы «Критические точки распределения Стьюдента»

По формуле (2) определяем границы доверительного интервала:

$$7,93 - 2,2 \frac{1,14}{\sqrt{12-1}} \leq \mu \leq 7,93 + 2,2 \frac{1,14}{\sqrt{12-1}},$$

$$7,93 - 0,76 \leq \mu \leq 7,93 + 0,76$$

$$7,17 \leq \mu \leq 8,69.$$

Таким образом, можно утверждать, что доверительный интервал (7,17; 8,69) (млн руб.) покрывает среднее значение объема заказов с вероятностью 0,95.

б) Найдем доверительный интервал для медианы.

Значение t_α мы уже нашли в предыдущем пункте: $t_\alpha = 2,2$. Далее используем формулу (3):

$$7,95 - 2,2 \sqrt{\frac{3,1415}{2}} \frac{1,14}{\sqrt{12-1}} \leq Me \leq 7,9 + 2,2 \sqrt{\frac{3,1415}{2}} \frac{1,14}{\sqrt{12-1}},$$

$$7,002 \leq Me \leq 8,9.$$

Доверительный интервал для медианы с надежностью 0,95 равен (7,0; 8,9) (млн руб.).

в) Найдем доверительный интервал для дисперсии.

По таблице «Критические точки распределения χ^2 » (приложение, таблица II) найдем квантили распределения χ^2 с $\nu = n - 1 = 12 - 1 = 11$ степенями свободы (рис. 3):

$$\chi^2_{\alpha} = \chi^2_{0,05} = 19,675 \text{ и } \chi^2_{1-\alpha} = \chi^2_{0,95} = 4,575.$$

Критические точки распределения χ^2

$\alpha \backslash \nu$	0,99	0,95	0,90	0,10	0,05	0,02	0,01	0,001
...								
10	2,558	3,940	4,865	15,987	18,307	21,161	23,209	29,588
11	3,053	4,575	5,578	17,275	19,675	22,618	24,725	31,264
12	3,571	5,226	6,304	18,549	21,026	24,054	26,217	32,909

Рис. 3. Фрагмент таблицы «Критические точки распределения χ^2 »

Вычислим доверительный интервал для дисперсии по формуле (4):

$$\frac{12 \cdot 1,3}{19,675} \leq \sigma^2 \leq \frac{12 \cdot 1,3}{4,575},$$

$$0,79 \leq \sigma^2 \leq 3,4.$$

Доверительный интервал для стандартного отклонения:

$$\sqrt{0,79} \leq \sigma \leq \sqrt{3,4} \quad \text{или} \quad 0,889 \leq \sigma \leq 1,844.$$

С надежностью 0,95 доверительный интервал для дисперсии равен $(0,79; 3,4)$ (млн руб.)², а для стандартного отклонения – $(0,889; 1,844)$ (млн руб.). ■

► **Пример 3.** Ежегодно в мае по всему миру проходит акция «Ночь музеев», участники которой могут посетить специально подготовленные музеями программы в вечернее и ночное время. Музеи города инициировали социологический опрос с целью выяснить, какое количество жителей города планирует принять участие в этой акции. Из 175 опрошенных 37 человек высказали желание посетить музеи в этот день/ночь. С вероятностью 0,99 оцените границы доли респондентов, которые собираются принять участие в этой акции.

Решение. Из условия известно, что $n = 175$, $m = 37$, $\gamma = 0,99$. Тогда уровень значимости равен $\alpha = 1 - \gamma = 0,01$. Границы доверительного интервала будем искать по формуле (5).

Вычислим относительную частоту: $w = \frac{m}{n} = \frac{37}{175} = 0,21$.

Найдем $u_{0,01}$ из уравнения $\Phi(u_{0,01}) = 1 - \frac{\alpha}{2} = 1 - \frac{0,01}{2} = 0,995$. В таблице «Нормальное распределение» (рис. 4, с.51) нет значения 0,995, а есть ближайшие значения 0,9949 и 0,9951. Поскольку 0,995 есть среднее арифметическое из 0,9949 и 0,9951, то $u_{0,01} = \frac{2,57 + 2,58}{2} = 2,575$.

Нормальное распределение $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$

x	Сотые доли				
	...	6	7	8	9
...		...	...	...	...
2,4		0,9931	0,9932	0,9934	0,9936
2,5		0,9948	0,9949	0,9951	0,9952
2,6		0,9961	0,9962	0,9963	0,9964

Рис. 4. Фрагмент таблицы «Нормальное распределение»

Следует заметить, что квантили с любой степенью точности можно найти с помощью статистических пакетов или, хотя бы, Excel.

Вычисляем доверительный интервал по формуле (5):

$$0,21 - 2,575 \sqrt{\frac{0,21(1-0,21)}{175}} \leq p \leq 0,21 + 2,575 \sqrt{\frac{0,21(1-0,21)}{175}},$$

$$0,0319 \leq p \leq 0,2909.$$

С вероятностью 0,99 можно утверждать, что доля жителей города, которые собираются принять участие в акции «Ночь музеев», заключена от 3,19 % до 29,09 %. ■

Контрольные вопросы

1. В каких случаях строятся доверительные интервалы?
2. Какие значения уровня значимости являются наиболее употребительными?
3. Доверительный интервал будет шире при уровне значимости 0,05 или 0,01?
4. Влияет ли выбор уровня значимости α на надежность оценки?

Задачи к §3.2

1. В страховой компании по результатам года был проведен анализ деятельности. В частности по полисам ОСАГО (27 шт.) в группе «легковой автомобиль с мощностью двигателя 71–100 л. с., выпуска 2013 г., 1 водитель – мужчина 30–35 лет с 7-летним водительским стажем, использующий автомобиль для личных нужд, проживающий в городе (от 500 тыс. чел.)» средние выплаты составили 49,2 тыс. руб., а выборочная дисперсия – $65,7$ (тыс. руб)². Постройте 95 %-е доверительные интервалы а) для среднего размера выплат; б) для дисперсии.

2. Исследование работы 56 лазерных принтеров показало, что среднее время безотказной работы равно 1500 ч. Найдите границы доверительных интервалов (вероятности 0,98 и 0,9) для среднего времени безотказной работы, если из априорных наблюдений известно, что стандартное отклонение равно 50 ч.

3. С целью оптимизации расходов было проведено исследование расходов отдела продаж за год. В частности, оказалось, что медианное значение расхода бумаги в месяц составило 5 пачек при выборочной дисперсии 14. Определите с вероятностью 0,95 границы медианного интервала.

4. 22 сентября – Международный день без автомобиля. На вопрос «Готовы ли вы отказаться от личного автомобиля?» из 162 респондентов 40 человек выбрали ответ «только в крайнем случае», а 47 опрошенных – «да, с удовольствием». Оцените с надежностью 0,96 долю респондентов, которые полагают, что а) могут отказаться от автомобиля только в крайнем случае; б) с удовольствием обойдутся без автомобиля.

5. Выборочное обследование бюджетов домашних хозяйств области по источникам поступления и потребления основных продуктов питания дало следующие результаты по потреблению картофеля.

Потребление картофеля

	Значение показателя, в среднем на потребителя, в год, кг	Выборочное стандартное отклонение	<i>n</i>
	в городской местности		
куплено	38,8	2,8	19
натуральные поступления	21,3	2,1	17
потреблено	60,5	1,7	26
	в сельской местности		
куплено	18,7	3,3	25
натуральные поступления	56,2	5,2	35
потреблено	71,8	2,7	38

1) Постройте доверительные интервалы для среднего значения объема потребления картофеля а) в городской местности; б) в сельской местности.

2) Постройте доверительные интервалы для среднего значения объема натурального поступления картофеля а) в городской местности; б) в сельской местности.

3) Постройте доверительные интервалы для стандартного отклонения объема купленного картофеля а) в городской местности; б) в сельской местности.

Надежность равна 0,99.

6. В результате выборочного обследования покупателей супермаркета получено следующее распределение по размеру сделанных покупок.

Стоимость покупок, руб.	До 100	100–200	200–300	300 и более
Число покупателей	17	58	89	43

С вероятностью 0,97 определите: а) границы среднего размера покупки; б) границы удельного веса (доли) покупок на сумму до 100 руб.

7. Вариация (стандартное отклонение) ежедневного дохода случайно выбранных 54 киосков оказалась равной 95 д.е. Найдите предельную ошибку выборки, которая с надежностью 0,98 определяет среднесуточный доход.

8. В универмаге планируется провести очередное изучение структуры и стоимости покупок. Сколько следует опросить покупателей, чтобы с вероятностью 0,95 определить стоимость покупки с точностью не менее 50 д.е. Дисперсию принять равной 62 500 (по прошлому обследованию).

9. Найти точность, с которой определен средний годовой привес скота (200 кг) для стада из 81 головы, если стандартное отклонение равно 48 кг. Оценку произвести, приняв доверительные вероятности $\gamma_1 = 0,9$; $\gamma_2 = 0,95$; $\gamma_3 = 0,99$. Как отличаются оценки?

10. Имеются следующие данные о дневном надое коров на молочно–товарной ферме.

Дневной надой, л.	13	14	15	16	17	18	19	20
Число коров	4	8	12	25	45	32	33	26

С вероятностью 0,99 постройте доверительные интервалы: а) для среднего надоя коров; б) для доли коров, имеющих надой выше среднего по данной ферме.

11. На диаграммах (рисунки 1, 2, с. 55) представлены результаты опроса на тему «Как часто Вы пользуетесь Интернетом?» с интервалом в 10 лет: октябрь 2006 г. и октябрь 2016 г. Респондентам были предложены следующие варианты ответов: а) практически ежедневно; б) иногда (несколько раз в неделю, в месяц); в) эпизодически, но не менее 1 раза в полгода; г) не пользуюсь; д) затрудняюсь ответить.

Как часто Вы пользуетесь Интернетом?

X.2006

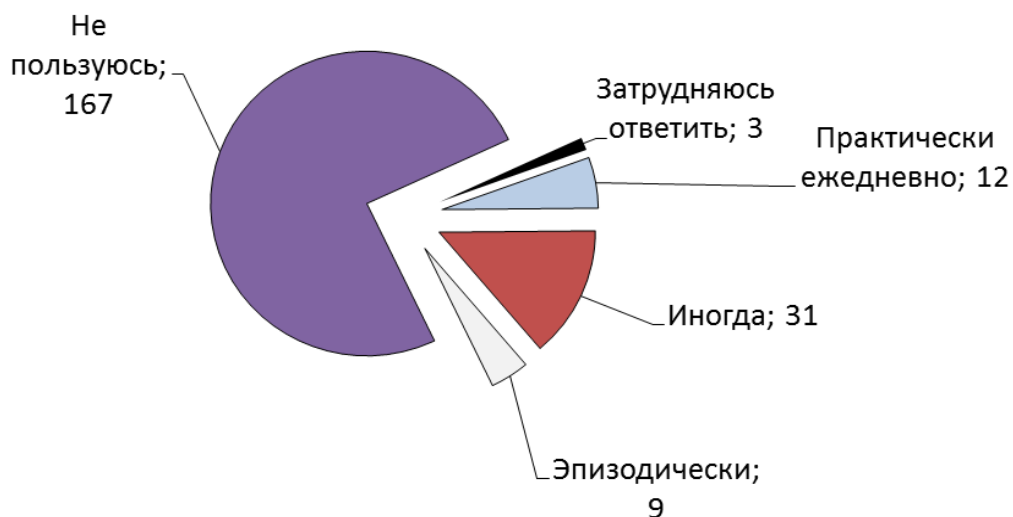


Рис. 1

X.2016

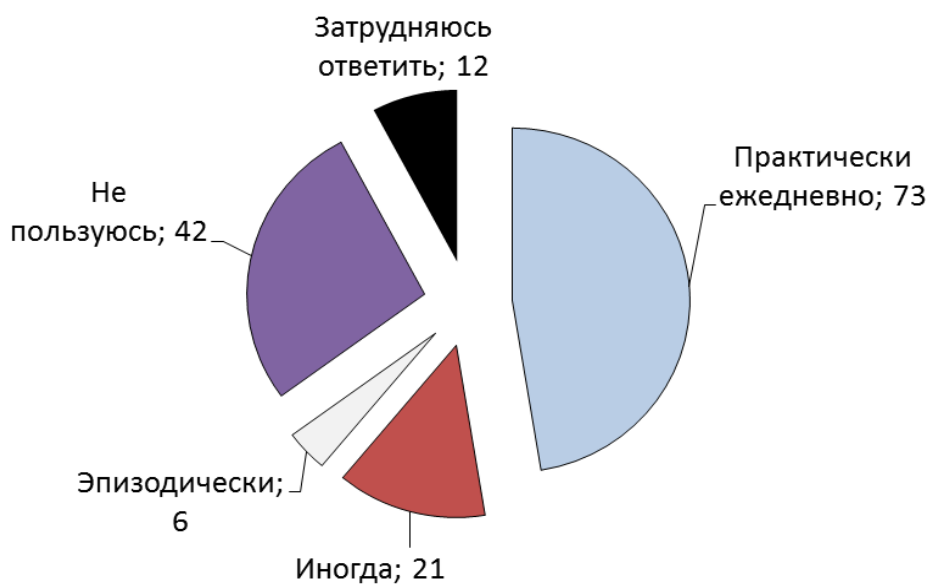


Рис. 2

Постройте доверительные интервалы с надежностью 0,98 для доли людей, которые практически ежедневно используют Интернет по данным 1) 2006 г.; 2) 2016 г.

4. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Ключевые слова: статистические гипотезы, нулевая и конкурирующая гипотезы, ошибки I и II рода, уровень значимости, критическая область.

На практике часто приходится по результатам обследований, испытаний, опытов проверять различные предположения о характеристиках исследуемого объекта. В этом случае решение об истинности предположения принимается путем проверки статистической гипотезы.

Статистической гипотезой, обозначаемой H , называется любое предположение относительно вида или параметра распределения случайной величины X .

Нулевой гипотезой называют проверяемую гипотезу H_0 .

Альтернативной (конкурирующей) называют гипотезу H_1 , которая противоречит нулевой.

Правило, по которому мы решаем принять или отклонить гипотезу H_0 , называется *критерием* K .

Так как решение мы будем принимать на основе наблюдений с.в. X , то нам необходимо выбрать подходящую статистику, которая называется *статистикой* T критерия K . Критерий, как статистическое правило, определяется заданием критической области – описанием тех данных, при которых мы отклоняем гипотезу H_0 .

Критической областью называют множество всех значений статистики критерия, при которых нулевая гипотеза отклоняется. Критической областью изображена на рис. 1–3 (с. 57–58).

Эта область обычно задается с помощью статистики T и имеет вид неравенства $T(c) > c$, где c выбирается по заданному уровню значимости.

Положение критической области зависит от формулировки альтернативной гипотезы H_1 . Предположим, что проверяется гипотеза $H_0: \theta = \theta_0$, а альтернативная гипотеза – $H_1: \theta \neq \theta_0$. Т.е. допускается, что различие может быть в обе стороны. Поэтому такие критерии называются *двусторонними*. На рис. 1 изображена двусторонняя критическая область

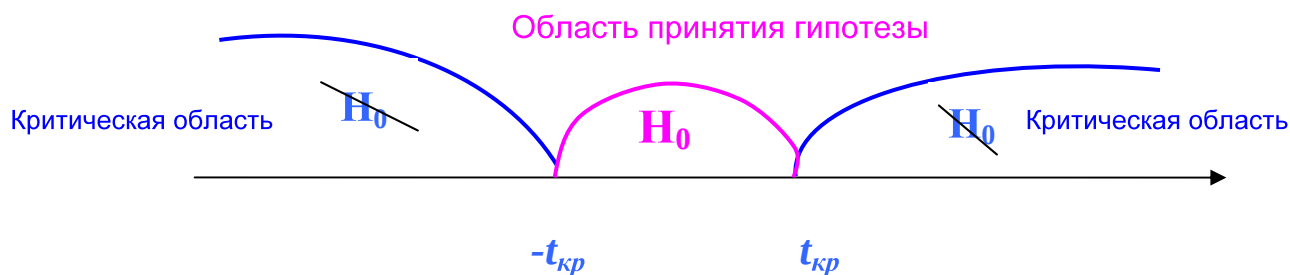


Рис. 1

В случае альтернативной гипотезы $H_1: \theta > \theta_0$ предполагается, что значение θ может быть расположено на числовой оси правее значения θ_0 . Поэтому критерий называется *односторонним*, а точнее – *правосторонним*. Его критическая область изображена на рис.2.

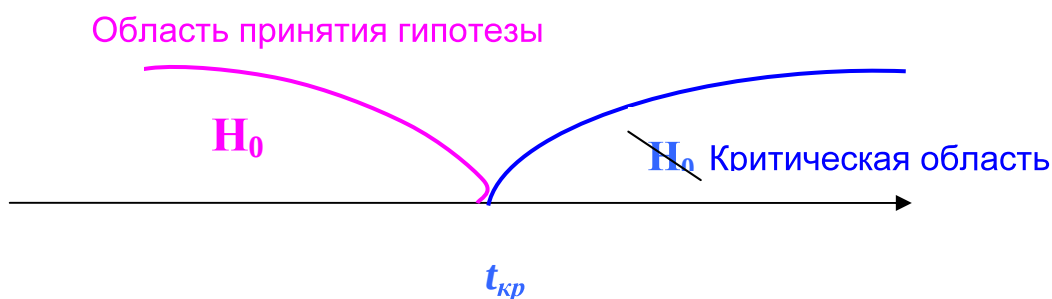


Рис. 2

Другим вариантом одностороннего критерия является левосторонний критерий: $H_1: \theta < \theta_0$. На рис. 3 (с. 58) изображена левосторонняя критическая область.

При проверке гипотез возможны два типа ошибок: 1) отклонение гипотезы H_0 , когда она верна – *ошибка первого рода*; 2) принятие

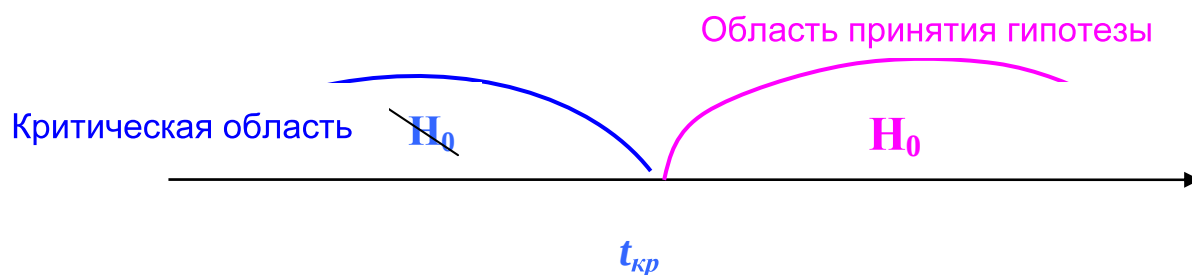


Рис. 3

гипотезы H_0 , когда она ложна – *ошибка второго рода*. Заданные ограничения на вероятность ошибки первого рода обозначаются α и называются *уровнем значимости критерия*. Вероятность ошибки второго рода обозначается β . Вероятности α и β удобно представить, как это сделано в следующей таблице.

Ошибки при проверке гипотез

	Решение	
	Принять H_0	Принять H_1
Справедлива H_0	Правильное с вероятностью $1 - \alpha$	Ошибочное с вероятностью α
Справедлива H_1	Ошибочное с вероятностью β	Правильное с вероятностью $1 - \beta$

Вероятность $(1 - \beta)$ не допустить ошибку 2-го рода, т.е. отвергнуть гипотезу H_0 , когда она неверна, называется *мощностью критерия*.

Следует особо подчеркнуть, что любая гипотеза должна формулироваться, уровень значимости α задаваться, а критерий выбираться исследователем всегда до получения экспериментальных данных, по которым эта гипотеза будет проверяться.

Проверка статистических гипотез может быть разбита на ряд этапов:

1. Сформулировать нулевую и альтернативную гипотезы.

Назначить уровень значимости α .

Выбрать критерий.

2. Получить экспериментальные данные.

3. Вычислить статистику (например, T) данного критерия.

4. Принять статистическое решение. В зависимости от имеющихся таблиц мы будем использовать один из следующих способов.

Способ 1. По таблице находим границы критической области.

1) Если значение статистики T попадает в область принятия гипотезы, то принимается гипотеза H_0 . То есть считается, что гипотеза H_0 не противоречит результатам наблюдений.

2) Если значение статистики попадает в критическую область, то гипотеза H_0 отклоняется как не согласующаяся с результатами наблюдений (рис.4).

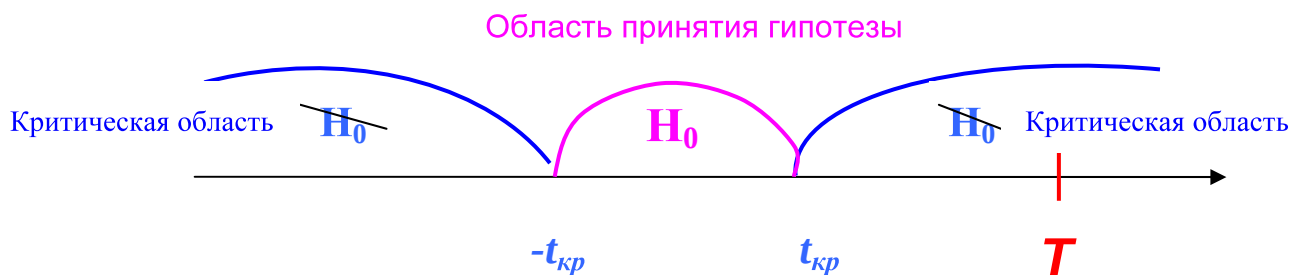


Рис. 4

Способ 2. По таблице находится критический уровень значимости $\alpha_{кр}$. Гипотеза отвергается, если критический уровень значимости $\alpha_{кр}$ меньше заданного α . Способ 2 обладает большей информативностью.

Контрольные вопросы

1. Как следует выбирать нулевую гипотезу?
2. Какие ошибки существуют при проверке гипотез?
3. Как определяется вероятность ошибки 1-го рода?
4. Какой уровень значимости лучше выбрать – 5% , 10% или 1%?

4.1. Проверка гипотезы о величине среднего значения нормального распределения (одновыборочный критерий Стьюдента)

Условия применения t-критерия: случайная выборка $X_1, \dots, X_n$ получена из нормального распределения с неизвестным средним μ и неизвестной дисперсией σ^2 .

Гипотеза $H_0: \mu = \mu_0$

Альтернатива $H_1: \mu \neq \mu_0$ или $H_1: \mu > \mu_0$ ($H_1: \mu < \mu_0$)

Уровень значимости: α .

Порядок применения:

1. Принимается предположение о нормальности, формулируется гипотеза H_0 и альтернатива H_1 , задается уровень значимости α .
2. Получают выборку $X_1, \dots, X_n$ объемом n .
3. Вычисляется статистика Стьюдента:

$$T = \frac{\bar{x} - \mu_0}{S} \sqrt{n-1},$$

где $\bar{x}, S_x^2$ – выборочные среднее и дисперсия выборки $X_1, \dots, X_n$.

Величина T при справедливости нулевой гипотезы имеет t -распределение Стьюдента с $\nu = n - 1$ степенями свободы.

4. Границы критической области вычисляются в зависимости от выбранной альтернативы.

а) Если $H_1 : \mu \neq \mu_0$, то $t_{кр} = t_{двуст\ кр}(\alpha; \nu)$. Значение $t_{кр} = t_{двуст\ кр}(\alpha; \nu)$ с ν степенями свободы находится по таблице «Критические точки распределения Стьюдента» (приложение, таблица III) по заданному уровню значимости α , помещенному в верхней строке таблицы (для двусторонней критической области). Если $|T| > t_{кр}$, то гипотеза H_0 отвергается.

б) Если $H_1 : \mu > \mu_0$, то $t_{кр} = t_{прав\ кр}(\alpha; \nu)$. Если $T > t_{кр}$, то гипотеза H_0 отвергается.

в) Если $H_1 : \mu < \mu_0$, то $t_{кр} = t_{лев\ кр} = -t_{пр\ кр}(\alpha; \nu)$. Если $T < t_{лев\ кр}$, то гипотеза H_0 отвергается.

Значения $t_{кр} = t_{прав\ кр}(\alpha; \nu)$, $t_{кр} = t_{лев\ кр} = -t_{пр\ кр}(\alpha; \nu)$ с ν степенями свободы находится по таблице «Критические точки распределения Стьюдента» по заданному уровню значимости α , помещенному в нижней строке таблицы (для односторонней критической области).

Вывод: если гипотеза H_0 отвергается, то выборочное среднее значительно отличается от μ_0 на уровне значимости α . В противном случае различие не является статистически значимым.

► **Пример.** Руководство компании приняло решение лишить бригаду премии по причине регулярного перерасхода материалов. Бригадир подобное утверждение посчитал несправедливым, аргументируя тем, что средний расход материала соответствует норме – 25 м. Для экспертизы было произведено 12 выборочных замеров. Расчеты показали, что средний расход материала равен 26,1 м, а выборочная дисперсия – 8,9. Какие выводы можно сделать по данным экспертизы при уровне значимости 0,05? Можно ли утверждать, что бригаду лишили премии обоснованно?

Решение. Используем t -критерий и действуем в указанном порядке.

1. Предполагаем, что распределение расхода материала приближенно нормальное.

Гипотеза $H_0: \mu = 25$.

Альтернатива $H_1: \mu \neq 25$.

Мы поставили альтернативную гипотезу о несоответствии норме. Также в качестве альтернативной гипотезы можно было взять одностороннюю (правостороннюю) гипотезу $H_1: \mu > 25$, т.к. в задаче речь идет о превышении нормы – перерасходе материала.

Уровень значимости: $\alpha = 0,05$.

2.-3. Из условия задачи известно: $n = 12$, $\bar{x} = 26,1$; $S_x^2 = 8,9$; $S_x = 2,98$.

$$t = \frac{\bar{x} - \mu_0}{S} \sqrt{n-1} = \frac{26,1 - 25}{2,98} \sqrt{12-1} = 1,22$$

$$\text{и } v = n - 1 = 12 - 1 = 11.$$

4. Находим по таблице «Критические точки распределения Стьюдента» по заданному уровню значимости $\alpha = 0,05$, помещенному в верхней строке таблицы (для двусторонней критической области), и $v = 11$ значение $t_{кр} = t_{\text{двуст.кр}}(\alpha; v) = t_{\text{двуст.кр}}(0,05; 11) = 2,2$. Т.е. критическая область есть $(-\infty; -2,2) \cup (2,2; +\infty)$. На рис. 1 изображена критическая область.

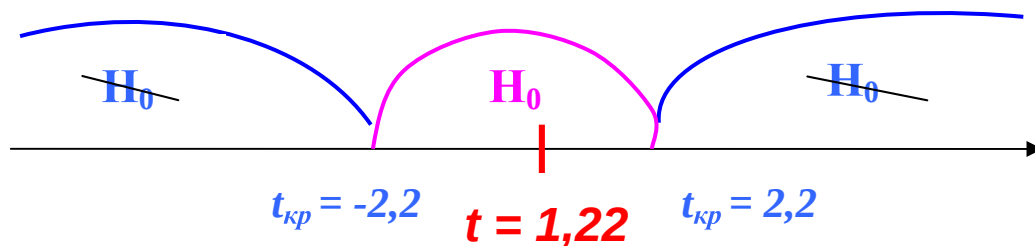


Рис. 1

Так как $|t| < t_{кр} (1,22 < 2,2)$, то гипотеза H_0 принимается.

Вывод: имеющиеся данные подтверждают, что средний расход материала соответствует норме. Следовательно, бригаду лишили премии необоснованно. ■

Контрольные вопросы.

1. Чему равна статистика Стьюдента?
2. Какое распределение имеет статистика Стьюдента?
3. Какое условие налагается на применение критерия Стьюдента?
4. В каких случаях применяются односторонние и двусторонние критерии?

Задачи к §4.1

В задачах данного раздела предполагается, что выборки имеют нормальное распределение, либо достаточно близкое к нормальному.

1. В инструкции по эксплуатации фена утверждается, что фен рассчитан на 400 часов работы. Для выборки из 26 фенов средний срок службы составил 380 ч. при выборочном стандартном отклонении 50 ч. Проверьте утверждение инструкции при уровне значимости 0,05.

2. По выборке объема $n = 21$, извлеченной из нормальной совокупности, были найдены выборочное среднее $\bar{x} = 17,5$ и выборочное стандартное отклонение $S = 3,5$. Необходимо при уровне значимости 0,05 проверить гипотезу $H_0 : \mu = 16$ при альтернативе а) $H_1 : \mu \neq 16$; б) $H_1 : \mu > 16$.

3. Ученики 7 класса должны решить тестовое задание за 30 мин. Индивидуальное время выполнения каждым учеником этого задания составило: 29, 31, 30, 27, 25, 34, 36, 35, 29, 30, 28, 42, 31, 29, 29, 32, 27 (мин.). Можно ли утверждать, что класс в среднем выполнил задание в указанное время при уровне значимости 0,05?

4. Часовая выработка забойщика при добыче угля в шахте по норме составляет 400 кг. Фактическая выработка соответствовала норме. При переходе в новый забой условия работы забойщиков усложнились. Поэтому была создана специальная комиссия с целью определить необходимость пересмотра нормы добычи угля. Для этого

был проведен учет работы 9 забойщиков. Их средняя часовая выработка составила 388 кг с выборочной дисперсией 171кг^2 . Какой вывод должна сделать комиссия, если уровень значимости равен 0,01?

5. Общество по защите прав потребителей проводило проверку качества продукции. Для этого было сделано 12 контрольных закупок банок кофе, масса нетто 95 г. фирмы “MN”. В результате средняя масса нетто составила 93 г., а выборочная дисперсия – 7,6. Проверить при уровне значимости 0,05, соответствует ли масса нетто норме, указанной на банке.

6. В период подготовки к экзамену студенты (группа девушек) были протестированы по тесту Спилберга на уровень реактивной тревожности. С целью снижения тревожности в группе был проведен тренинг по новой методике и студентки еще раз были протестированы по тесту Спилберга (в период подготовки к экзамену). Результаты представлены в следующей таблице.

Уровень реактивной тревожности	Испытуемые (код имени)							
	И.А.	К.Л.	В.Д.	К.О.	Н.В.	В.З.	Ф.О.	Ю.Ю.
До тренинга	32	34	28	43	35	26	41	32
После тренинга	26	29	31	35	29	26	31	24

Подтверждают ли результаты тестирования эффективность новой методики? Проверьте это предположение на уровне значимости 0,05.

(Указание. Предлагается создать новую выборку $z_i = x_i - y_i$, где x_i и y_i – уровни тревожности у девушки до и после тренинга соответственно. Далее проверить гипотезу $H_0 : \mu_z = 0$, которая означает, что никаких статистически значимых изменений не произошло.)

7. Как изменится статистика критерия Стьюдента, если вместо выборочной дисперсии мы подставим несмещенную выборочную дисперсию?

4.2. Проверка гипотезы о равенстве средних значений двух нормальных распределений (двухвыборочный критерий Стьюдента)

Условия применения t-критерия: случайные выборки $X_1, \dots, X_n$ и $Y_1, \dots, Y_m$ имеют нормальное распределение с неизвестными средними μ_x и μ_y и общей неизвестной дисперсией σ^2 ; $n, m < 30$.

Гипотеза $H_0 : \mu_x = \mu_y$.

Альтернатива $H_1 : \mu_x \neq \mu_y$ или $H_1 : \mu_x > \mu_y$ ($\mu_x < \mu_y$) в зависимости от того, что требуется доказать: простое различие средних или то, что одно из них больше другого.

Уровень значимости: α .

Порядок применения:

1. Принимается предположение о нормальности, формулируется гипотеза H_0 и альтернатива H_1 , задается уровень значимости α .
2. Получают две независимые выборки $X_1, \dots, X_n$ и $Y_1, \dots, Y_m$ объемами n и m соответственно.
3. Вычисляется статистика Стьюдента:

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{nS_x^2 + mS_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}},$$

где $\bar{x}, S_x^2$ – выборочные среднее и дисперсия выборки $X_1, \dots, X_n$,

а $\bar{y}, S_y^2$ – второй выборки $Y_1, \dots, Y_m$.

Величина T при справедливости нулевой гипотезы имеет t -распределение Стьюдента с $\nu = n + m - 2$ степенями свободы.

4. Границы критической области вычисляются в зависимости от выбранной альтернативы.

а) $H_1 : \mu_x \neq \mu_y$ $t_{кр} = t_{двуст\ кр}(\alpha; \nu)$. Значение $t_{кр} = t_{двуст\ кр}(\alpha; \nu)$ с ν степенями свободы находится по таблице «Критические точки распределения Стьюдента» по заданному уровню значимости α , помещенному в верхней строке таблицы (для двусторонней критической области). Если $|T| > t_{кр}$, то гипотеза H_0 отвергается.

б) $H_1 : \mu_x > \mu_y$ $t_{кр} = t_{прав\ кр}(\alpha; \nu)$. Если $T > t_{кр}$, то гипотеза H_0 отвергается.

в) $H_1 : \mu_x < \mu_y$ $t_{кр} = t_{лев\ кр} = -t_{пр\ кр}(\alpha; \nu)$. Если $T < t_{лев\ кр}$, то гипотеза H_0 отвергается.

Значения $t_{кр} = t_{прав\ кр}(\alpha; \nu)$, $t_{кр} = t_{лев\ кр} = -t_{пр\ кр}(\alpha; \nu)$ с ν степенями свободы находится по таблице «Критические точки распределения Стьюдента» по заданному уровню значимости α , помещенному в нижней строке таблицы (для односторонней критической области).

Вывод: если гипотеза H_0 отвергается, то выборочные средние значительно различаются на уровне значимости α . В противном случае различие не является статистически значимым.

 **Пример.** Опрос 22 мужчин и 20 женщин в возрасте 30–40 лет выявил, что среднее время поиска работы для безработного мужчины составляет 6,5 месяцев, а для женщин – 9,8 месяцев. Выборочные дисперсии времени поиска работы для мужчин равны 5,5; для женщин – 8,9. Подтверждают ли результаты исследования известное выражение «У безработицы женское лицо»? Принять уровень значимости 0,01.

Решение. Справедливость выражения «У безработицы женское лицо» в контексте наших данных означает, что мужчины значительно быстрее находят работу, чем женщины. Поэтому среди зарегистрированных безработных обычно преобладают женщины.

Обозначим с.в. X – время поиска работы для женщин;

с.в. Y – время поиска работы для мужчин.

Используем t -критерий и действуем в указанном порядке.

1. Предполагаем, что время поиска работы является приближенно нормальным.

Гипотеза $H_0 : \mu_x = \mu_y$.

Альтернатива $H_1 : \mu_x > \mu_y$.

Уровень значимости: $\alpha = 0,01$.

2. Из условия задачи известно:

$$\bar{x} = 9,8; \quad S_x^2 = 8,9; \quad n = 20,$$

$$\bar{y} = 6,5; \quad S_y^2 = 5,5; \quad m = 22.$$

3. Вычисляем статистику T :

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{nS_x^2 + mS_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}} = \frac{9,8 - 6,5}{\sqrt{20 \cdot 8,9 + 22 \cdot 5,5}} \sqrt{\frac{20 \cdot 22(20 + 22 - 2)}{20 + 22}} = 3,9,$$

и $v = 20 + 22 - 2 = 40$.

4. Находим $t_{кр}$ по таблице «Критические точки распределения Стьюдента» по заданному уровню значимости $\alpha = 0,01$, помещенному в нижней строке таблицы (для односторонней критической области):

$$t_{кр} = t_{прав\ кр}(\alpha; v) = t_{прав\ кр}(0,01; 40) = 2,42.$$

Критическая область имеет следующий вид.

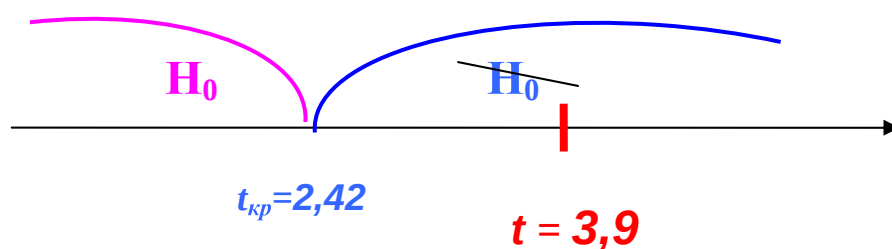


Рис. 1

Поскольку $t > t_{кр}$ ($3,9 > 2,42$), то гипотеза H_0 отвергается.

Вывод: на основе имеющихся данных можно сделать вывод, что мужчины значительно быстрее находят работу, чем женщины. Таким образом, подтверждается справедливость выражения «У безработицы женское лицо» ■

Контрольные вопросы

1. Можно ли применять критерий для двух зависимых выборок?
2. В каких случаях применяются односторонние и двусторонние критерии?
3. Как могут быть проверены гипотезы об однородности, репрезентативности выборки?
4. Какое распределение имеет статистика Стьюдента?

Задачи к §4.2

В задачах данного раздела предполагается, что выборки имеют нормальное распределение, либо достаточно близкое к нормальному.

1. В результате исследования продолжительности простоев рабочих (12 токарей и 10 ткачих) по организационно-техническим причинам определили, что средние простои у токарей равны 20,5 мин., у ткачих – 18,5 мин., а выборочные стандартные отклонения, соответственно – 0,7 и 0,8. Выяснить при уровне значимости 0,01 существенность различия средних простоев токарей и ткачих.

2. По двум независимым выборкам объема $n = 11$ и $m = 18$, извлеченных из нормальной совокупности, были найдены выборочные средние $\bar{x} = 42$, $\bar{y} = 45$ и выборочные стандартные отклонения $S_x = 3,78$ и $S_y = 2,91$. Необходимо проверить гипотезу $H_0 : \mu_x = \mu_y$ при альтернативе $H_1 : \mu_x \neq \mu_y$ на уровне значимости а) 0,05; б) 0,01.

3. Сколько нужно денег для счастья? Молодые люди (18–25 лет) считают, что для счастья хватит 98,4 тыс. руб. в месяц (в среднем), а для людей среднего возраста (30–45 лет) среднее значение составило 78 тыс. руб. Значимо ли отличаются представления о денежной базе для счастья у людей разных возрастов? Проверьте это предположение на уровне значимости 0,01. В исследовании участвовали 24 человека молодого возраста и 19 человек – среднего возраста. Выборочные стандартные отклонения соответственно равны 12,2 и 18,4 (тыс. руб.)

4. Принцип Принса (один из следствий закона Мерфи) гласит: «Людам, которые работают сидя, платят больше, чем людям, которые работают стоя». Для проверки этого принципа были взяты респонденты из телефонного опроса. В итоге были получены следующие данные.

Зарплата (тыс. д.е.)			
Люди, которые работают сидя		Люди, которые работают стоя	
35	34	16	28
45	28	32	41
28	17	27	34
19	54	19	27
47	27	24	39
18	-	14	17
41	-	16	18
14	-	29	-

Проверьте, подтверждают ли приведенные выше данные Принцип Принса.

5. Средненедельный объем продаж для 15 торговых точек района А составил 3100 тыс. д.е. при стандартном отклонении 400 тыс.д.е., а для 10 торговых точек района В – 2600 тыс. д.е. при стандартном отклонении 500 тыс. д.е. Значимо ли различие средненедельных объемов продаж в районах А и В при 5 %-м уровне значимости? Значимо ли превышение средненедельного объема продаж в районе А по сравнению с районом В? Чем отличается этот вопрос от предыдущего?

6. Согласно данным HR-агентств работодатели предпочитают брать на работу соискателей уже имеющих стаж работы не менее 5 лет в их сфере деятельности. Выборочный опрос нашедших работу соискателей дал следующие результаты.

№	Рабочий стаж, лет	Продолжительность поиска работы, мес.	№	Рабочий стаж, лет	Продолжительность поиска работы, мес.
1	5	2	13	6	3
2	1	5	14	1	4
3	8	8	15	9	5
4	3	12	16	1	1
5	5	9	17	7	1
6	2	8	18	4	7
7	12	3	19	8	1
8	2	7	20	2	13
9	4	14	21	8	4
10	3	9	22	3	8
11	7	5	23	6	8
12	4	2	24	2	9

Подтверждают ли полученные данные выводы HR-агентств?

(Указание. Из представленных наблюдений сформируйте 2 выборки. В выборку №1 (маленький стаж) включаются данные, соответствующие рабочему стажу менее 5 лет, а в выборку №2 (большой стаж) – не менее 5 лет.)

7. Для сравнения предлагаемого нового сорта пшеницы В со старым А под каждый сорт было отведено по 3 участка. В результате был получен урожай, в пересчете в центнерах на гектар, равный:

Сорт А – 15,6; 18,2; 13,3.

Сорт В – 17,1; 16,3; 18,8.

Определить, можно ли, допуская риск 5%, считать, что сорт В урожайнее сорта А?

4.3. Проверка гипотезы о равенстве дисперсий двух нормальных распределений (F –критерий Фишера)

Условия применения F –критерия: $X_1, \dots, X_n$ и $Y_1, \dots, Y_m$ – две независимые случайные выборки ($n, m < 100$) из нормальных распределений с дисперсиями σ_x^2 и σ_y^2 и неизвестными средними μ_x и μ_y .

Гипотеза $H_0 : \sigma_x^2 = \sigma_y^2$.

Альтернатива $H_1 : \sigma_x^2 \neq \sigma_y^2$. Если предположить, что одна из выборок имеет большую дисперсию (обозначим ее σ_1^2), чем другая (σ_2^2), то можно сформулировать одностороннюю гипотезу $H_1 : \sigma_1^2 > \sigma_2^2$.

Уровень значимости: α .

Порядок применения:

1. Принимается предположение о нормальности распределения выборок, формулируется гипотеза H_0 и альтернатива H_1 , задается уровень значимости α .
2. Получают две независимые выборки $X_1, \dots, X_n$ и $Y_1, \dots, Y_m$ объемами n и m соответственно.
3. Рассчитывают значения несмещенных выборочных дисперсий S_x^2 и S_y^2 . Большую из дисперсий (S_x^2 или S_y^2) обозначают S_1^2 , а меньшую дисперсию – S_2^2 .

Вычисляется значение F –критерия по формуле

$$F = \frac{S_1^2}{S_2^2},$$

$\nu_1 = n_1 - 1$ и $\nu_2 = n_2 - 1$, где ν_1 – число степеней свободы большей дисперсии S_1^2 .

4. По таблице «Процентные точки распределения Фишера-Снедекора (F-распределения)» (приложение, таблица IV) находят критическое значение F -критерия при заданном уровне значимости α и числе степеней свободы $\nu_1 = n_1 - 1$ и $\nu_2 = n_2 - 1$.

При альтернативе $H_1: \sigma_x^2 \neq \sigma_y^2$ критическую точку $F_{кр}(\frac{\alpha}{2}, \nu_1, \nu_2)$ находят по уровню значимости $\frac{\alpha}{2}$ и числу степеней свободы ν_1, ν_2 .

При альтернативе $H_1: \sigma_x^2 > \sigma_y^2$ находят критическую точку $F_{кр}(\alpha, \nu_1, \nu_2)$.

Если $F < F_{кр}$, то нет оснований отвергнуть нулевую гипотезу. То есть, различие дисперсий не является статистически значимым.

► **Пример.** По рекомендации психологов в бригаде №2 (экспериментальная группа) были введены дополнительные перерывы каждые 2 часа работы: 1-й перерыв между началом работы и обедом, 2-й перерыв – между обедом и окончанием работы. Во время перерыва работники отдыхали по специальной методике. Бригада №1 (контрольная группа) работала по старому графику. Выборочное исследование показало следующее время выполнения операций.

Бригада № 1: 10,1; 10,12; 10,2; 9,7; 10,4; 10,16; 10,18 (мин.).

Бригада №2: 10,14; 11; 11,2; 10,9; 10,8; 10,3 (мин.).

Можно ли утверждать, что экспериментальная группа имеет большую стабильность в работе, чем контрольная группа? Проверьте это предположение на уровне значимости 0,05.

Решение. В условиях нашей задачи выражение «большая стабильность» подразумевает, что разброс замеров времени выполнения операций меньше в экспериментальной группе по сравнению с контрольной.

Обозначим

с. в. X – значения замеров времени выполнения операций бригадой № 1,

с. в. Y – значения замеров времени выполнения операций бригадой № 2.

1. Гипотеза $H_0: \sigma_x^2 = \sigma_y^2$.

Альтернатива $H_1: \sigma_x^2 > \sigma_y^2$, т.к., согласно предположению разброс значений замеров в экспериментальной группе (Y) меньше, чем в контрольной (X).

Уровень значимости: $\alpha = 0,05$.

Принимаем предположение о нормальности распределения обеих выборок.

2.-3. По формулам (§§2.2.1, 2.2.2) вычисляем выборочные дисперсии. Исправленные (несмещенные) дисперсии найдем по формуле (§3.1.1):

$$S_{\text{несм}}^2 = \frac{n}{n-1} S^2.$$

Бригада №1 (контрольная группа): $n_x = 6$, $S_x^2 = 0,1434$,

$$S_{x,\text{несм}}^2 = \frac{6}{6-1} 0,1434 = 0,172,$$

Бригада №2 (экспериментальная группа): $n_y = 7$, $S_y^2 = 0,0381$,

$$S_{y,\text{несм}}^2 = \frac{7}{7-1} 0,0381 = 0,0445.$$

Обозначим

$$S_1^2 = \max(S_{x,\text{несм}}^2, S_{y,\text{несм}}^2) = \max(0,172; 0,0445) = 0,172 = S_{x,\text{несм}}^2.$$

Тогда $S_2^2 = S_{y,\text{несм}}^2 = 0,0445$.

Вычисляем значение F -критерия

$$f = \frac{S_1^2}{S_2^2} = \frac{0,172}{0,0445} = 3,87,$$

$$\nu_1 = n_1 - 1 = n_x - 1 = 6 - 1 = 5 \quad \text{и} \quad \nu_2 = n_2 - 1 = n_y - 1 = 7 - 1 = 6.$$

4. Найдем границы критической области по таблице «Процентные точки распределения Фишера-Снедекора (F-распределение)» (при-

ложение, таблица IV). Так как $\alpha = 0,05$, то используем таблицу «F-распределение: верхние 5%-ные точки» и находим $F_{кр}(\alpha; \nu_1; \nu_2) = F_{кр}(0,05; 5; 6) = 4,39$. Критическая область изображена на рис.1.

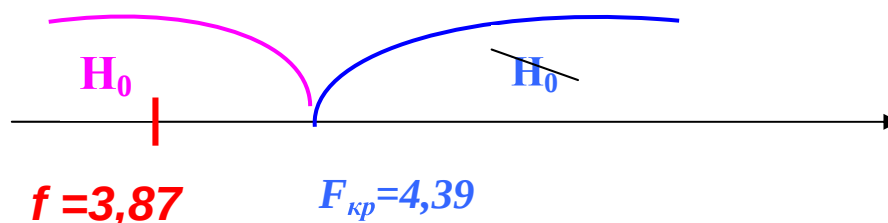


Рис. 1

Поскольку $f < F_{кр}$ ($3,87 < 4,39$), то гипотеза H_0 принимается. Следовательно, на уровне значимости 0,05 различие дисперсий не является статистически значимым.

Вывод: полученные данные не подтверждают предположения о том, что экспериментальная группа имеет большую стабильность в работе по сравнению с контрольной группой. Таким образом, эффективность новой методики, предложенной психологами, не подтверждается. ■

Контрольные вопросы

1. Какое условие налагается на применение критерия Фишера?
2. Чему равна статистика Фишера?
3. Какое распределение имеет статистика Фишера?
4. Можно ли применять критерий Фишера как предварительный тест для проверки условий применимости критериев Стьюдента?
5. Найдите по таблице критическое значение для статистики Фишера при уровне значимости 0,01 и объемах выборок 7 и 9.
6. Найдите по таблице критическое значение для статистики Фишера при объемах выборок 7 и 9 и уровне значимости а) 0,01; б) 0,05. Сравните полученные значения.

Задачи к §4.3

В задачах данного раздела предполагается, что выборки имеют нормальное распределение, либо достаточно близкое к нормальному.

1. Проведено исследование розничного товарооборота продовольственных магазинов в двух районах области, в первом районе 42 магазина, во втором – 25. Полученные в результате оценки стандартных отклонений в первом и втором районах соответственно равны 6,72 и 7,79 (тыс. руб.). Можно ли считать, что разброс розничного товарооборота магазинов районов неодинаков при уровне значимости 0,10?

2. По двум независимым выборкам объема $n = 13$ и $m = 16$, извлеченных из нормальной совокупности, были найдены выборочные дисперсии $S_x^2 = 25,6$ и $S_y^2 = 8,94$. Необходимо проверить гипотезу $H_0 : \sigma_x^2 = \sigma_y^2$ при альтернативе $H_1 : \sigma_x^2 > \sigma_y^2$ на уровне значимости а) 0,05; б) 0,01.

3. Для исследования двух типов удобрений на урожайность пшеницы было засеяно по 15 опытных участков. Несмещенные выборочные дисперсии, характеризующие вариацию урожайности, соответственно, равны 0,25 и 0,49. Проверить, зависит ли вариация урожайности пшеницы от типа внесенных удобрений а) при $\alpha = 0,1$; б) при $\alpha = 0,02$.

4. На двух токарных станках обрабатывается втулка. Выборочно были взяты 12 втулок, обработанных на первом станке, и 13 втулок – на втором. По данным этих выборок рассчитали выборочные стандартные отклонения: 0,76 (для первого станка), 0,52 (для второго станка). При уровне значимости 0,10 проверить гипотезу о том, что станки обладают одинаковой точностью.

5. Определить, можно ли считать существенно различными режимы работ трактористов А и В, если у тракториста А при средней глубине вспашки в 21 см выборочное стандартное отклонение, определенное в результате 30 замеров, равнялось 2 см, а у тракториста В

при той же средней глубине вспашки, определенное в результате 20 замеров, оказалось 3,9 см. Уровень значимости взять равным 0,10.

6. Для эксперимента были выбраны 2 группы учащихся, одна из которых имеет гуманитарную специализацию, а другая – по химии. Ниже представлены результаты контрольной работы по математике, максимальный балл равен 20.

Направление специализации	Оценка за контрольную работу, балл								
гуманитарное	8	15	18	9	12	17	6	11	17
химическое	15	12	17	9	10	18	-	-	-

Значения средних баллов по обоим направлениям специализации не имеют статистически значимых различий. Можно ли сказать, что и по вариативности результатов контрольной работы учащиеся обоих направлений специализации также не отличаются?

7. Как известно, вкусовые привычки, предпочтения, сформированные в детстве, особенно в семье являются достаточно устойчивыми и очень часто сохраняются на протяжении всей жизни. Проверим это утверждение на примере потребления рыбы. Согласно рекомендации Минздрава (Приказ Министерства здравоохранения и социального развития РФ от 2 августа 2010 г. N 593н «Об утверждении рекомендаций по рациональным нормам потребления пищевых продуктов, отвечающим современным требованиям здорового питания») норма потребления рыбы в неделю составляет примерно 300 г. для взрослого человека со средним весом, а в год – 18–22 кг на человека. Для исследования были выбраны женщины среднего возраста со средним весом. Выбор женщин обусловлен тем обстоятельством, что обычно именно женщины занимаются приготовлением пищи и составлением меню. В группу №1 включены женщины, у которых в детском возрасте рыбные блюда присутствовали в рационе не менее 2-х раз в неделю. У женщин группы №2 в детском возрасте потребление рыбы было несистема-

тичным, время от времени. Наше предположение состоит в том, что сложившиеся в детстве традиции потребления рыбы сохраняются и во взрослом возрасте. Поэтому и разброс в объемах потребления рыбы будет меньше в группе №1. Проверьте это предположение на уровне значимости 0,01.

	Потребление рыбы в неделю, г.								
Группа №1	290	320	150	310	270	170	260	-	-
Группа №2	100	295	360	50	400	280	260	0	380

8. Одни и те же изделия получают на двух производственных линиях. Качество изделия зависит от продолжительности процесса обработки сырья. На второй линии введены некоторые усовершенствования, сократившие вариацию времени обработки, в связи с чем качество изделий повысилось. Затем были проведены выборочные измерения вариации времени обработки на обеих линиях. В результате были получены следующие значения несмещенных выборочных дисперсий: $S_x^2 = 3.579_{мин}^2$ при 21 наблюдении и $S_y^2 = 2,5963_{мин}^2$ при 11 наблюдениях. Можно ли считать существенными расхождения между вариациями продолжительности процесса обработки сырья на первой и второй линиях? Проверить гипотезу при уровне значимости 0,10.

9. Компания решила инвестировать средства в ценные бумаги компаний А и В. Выборка из 21 наблюдения по активу А показала, что выборочная дисперсия ее доходности равна 0,072, а по активу В – 0,068 при 16 наблюдениях. Эксперты предполагают, что бумаги компании А менее устойчивы к колебаниям рынка, поэтому инвестировать средства следует только в бумаги компании В. Проверьте предположение экспертов на уровне значимости 0,05.

4.4. Проверка гипотезы о равенстве долей признака в двух совокупностях

Условия применения t-критерия: две независимые случайные выборки объемами n_1 и n_2 . Событие A в выборке №1 появляется с неизвестной вероятностью p_1 , а в выборке №2 – с неизвестной вероятностью p_2 .

Чаще всего берут исходную выборку и делят ее на подвыборки. Например, по полу, возрасту, образованию, семейному положению, месту проживания и т.д.

Гипотеза $H_0: p_1 = p_2$

Альтернатива $H_1: p_1 \neq p_2$ или $H_1: p_1 > p_2$ ($H_1: p_1 < p_2$)

Уровень значимости: α .

Порядок применения:

1. Формулируется гипотеза H_0 и альтернатива H_1 , задается уровень значимости α .
2. Получают две независимые случайные выборки: n_i – объем i -й выборки; m_i – количество единиц в i -й выборке, обладающих исследуемым свойством; $i = 1, 2$.
3. Вычисляется статистика:

$$Z = \frac{w_1 - w_2}{\sqrt{\bar{p} \cdot \bar{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}},$$

где $\bar{p} = \frac{m_1 + m_2}{n_1 + n_2}$; $\bar{q} = 1 - \bar{p}$.

4. Границы критической области вычисляются в зависимости от выбранной альтернативы.

а) Если $H_1 : p_1 \neq p_2$, то значение $z_{кр}$ находим из уравнения $\Phi(z_{кр}) = 1 - \frac{\alpha}{2}$. Если $|Z| > z_{кр}$, то гипотеза H_0 отвергается.

б) Если $H_1 : p_1 > p_2$, то значение $z_{кр}$ находим из уравнения $\Phi(z_{кр}) = 1 - \alpha$. Если $Z > z_{кр}$, то гипотеза H_0 отвергается.

в) Если $H_1 : p_1 < p_2$, то значение $z_{кр}$ находим из уравнения $\Phi(z_{кр}^*) = 1 - \alpha$ и $z_{кр} = -z_{кр}^*$. Если $Z < -z_{кр}$, то гипотеза H_0 отвергается.

Если гипотеза H_0 отклоняется, то различие долей является статистически значимым.

 **Пример.** 2 июня – Всемирный день правильного питания.

Россияне чаще всего разделяют понятия «здоровое» и «правильное» питание. Под здоровым питанием обычно понимают экологически чистые продукты, выращенные своими руками. А правильное питание – это сбалансированное питание, потребление продуктов, полезных для здоровья, режим питания и т.д.

Следите ли Вы за своим питанием?

Варианты ответа	Мужчины	Женщины
Да, соблюдаю диету, которую рекомендовал врач	3	14
Да, соблюдаю самостоятельно выбранную диету	9	18
Да, стараюсь есть здоровую пищу	35	43
Ем, что хочу, здоровье меня не беспокоит	38	22
Думать о качестве пищи нет возможности; ем, что придется	21	23
Затрудняюсь ответить	5	3

(Варианты ответов взяты из аналогичного ежегодного опроса ВЦИОМ).

Можно ли утверждать, что среди женщин больше сторонников внимательного отношения к питанию по сравнению с мужчинами. Проверьте это предположение на уровне значимости 0,01.

Решение. Нам нужно проверить статистическую гипотезу о том, что доля женщин, которые следят за своим питанием, превышает долю мужчин, которые также внимательно относятся к своему питанию. Будем считать, что выборка №1 – это женщины, соответственно выборка №2 – мужчины.

1. Гипотеза $H_0: p_1 = p_2$

Альтернатива $H_1: p_1 > p_2$

Уровень значимости: $\alpha = 0,01$.

2. Анализ таблицы (с. 79) позволяет сделать следующие выводы. В опросе приняли участие 123 женщины и 111 мужчин. Нас интересуют респонденты, которые (по разным причинам) следят за своим питанием. Так, среди женщин следят за своим питанием $14 + 18 + 43 = 75$ человек, а среди мужчин – 47 человек ($3 + 9 + 35 = 47$).

Итак,  женщины: $n_1 = 123$; $m_1 = 75$; $w_1 = \frac{m_1}{n_1} = \frac{75}{123} = 0,61$.

 мужчины: $n_2 = 111$; $m_2 = 47$; $w_2 = \frac{m_2}{n_2} = \frac{47}{111} = 0,423$.

3. Вычислим долю людей, которые следят за своим питанием, среди всех опрошенных: $\bar{p} = \frac{m_1 + m_2}{n_1 + n_2} = \frac{75 + 47}{123 + 111} = 0,521$.

Тогда $\bar{q} = 1 - \bar{p} = 1 - 0,521 = 0,479$.

Статистика $z = \frac{w_1 - w_2}{\sqrt{\bar{p} \cdot \bar{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{0,61 - 0,423}{\sqrt{0,521 \cdot 0,479 \left(\frac{1}{123} + \frac{1}{111} \right)}} = 2,85$.

4. Найдем границы критической области $z_{кр}$ из уравнения $\Phi(z_{кр}) = 1 - \alpha = 1 - 0,01 = 0,99$. В Примере 1, §3.2 (с. 47) подробно описано, как определяется квантиль нормального распределения. Получаем $z_{кр} = 2,33$. Односторонняя критическая область изображена на рис.1.

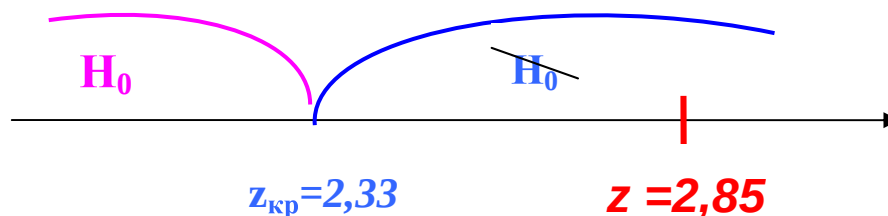


Рис. 1

Поскольку $\mathbf{z} > \mathbf{z}_{кр}$ ($2,85 > 2,33$), то гипотеза H_0 отклоняется. Следовательно, на уровне значимости 0,05 различие долей является статистически значимым.

Вывод: полученные данные подтверждают предположения о том, среди женщин больше сторонников внимательного отношения к питанию по сравнению с мужчинами. ■

Контрольные вопросы

1. Какие данные могут быть использованы в критерии?
2. Чему равна статистика критерия?
3. Какое распределение имеет статистика критерия?
4. Что означают величины $\bar{p}$ и $\bar{q}$?
5. Найдите по таблице критическое значение для статистики критерия при уровне значимости а) 0,03; б) 0,07.
6. В чем различие односторонней и двусторонней гипотез этого критерия?

Задачи к §4.4

1. С целью выявления источника информации для клиентов менеджеры по рекламе опросили клиентов компании о том, как они узнали об их компании. Так, в марте из 165 клиентов узнали о компании по Интернету 58 человек, а из бумажных СМИ – 35. Остальные – из других источников. В августе было 187 клиентов, из них 92 человек о компании узнали по Интернету, а из бумажных СМИ – 45. Свидетельствуют ли приведенные данные о том, что доля клиентов, получивших информацию о компании из Интернета, имеет тенденцию к увеличению?

2. В двух независимых выборках объемами $n_1 = 123$ и $n_2 = 111$ найдено число появлений исследуемого события A $m_1 = 75$ и $m_2 = 4$ соответственно. На уровне значимости 0,05 проверьте нулевую гипотезу о равенстве долей $H_0: p_1 = p_2$ при конкурирующей $H_1: p_1 \neq p_2$.

3. В детских садах проводилось исследование пространственного восприятия детей. В частности, детям была предложена иллюзия (рис. 1), которая была впервые описана немецким психиатром Францем Мюллером-Лайером в 1889 году.

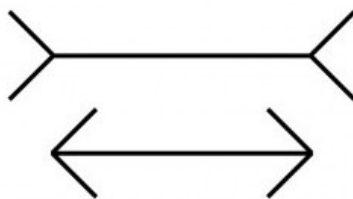


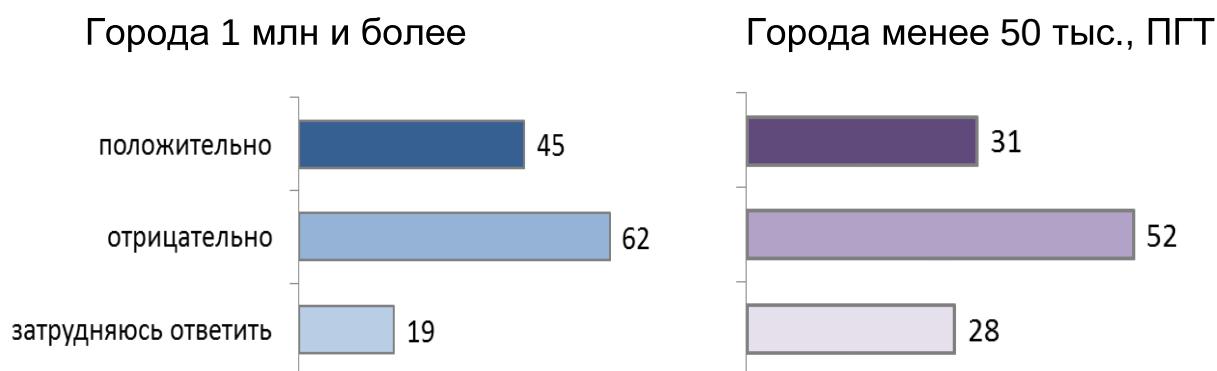
Рис. 1. Иллюзия Мюллера-Лайера

В группе 5-летних детей 35 человек из 91 ответили, что отрезки имеют одинаковую длину. В группе 6-летних детей так считают 57 из 121 человек. Свидетельствуют ли результаты этого исследования о том, что процент детей, имеющих адекватное пространственное восприятие, увеличивается к 6-летнему возрасту?

4. Для изучения электоральных настроений было проведено исследование (88 опрошенных), согласно которому в регионе политическую партию А поддерживает четверть опрошенного населения. Можно ли сказать, что в столице региона г. Речинск партия А имеет больше влияние, чем в целом по региону, если из 120 опрошенных жителей города Речинск приверженцами данной партии являются 45 человек.

5.

Если бы там, где вы живёте, ввели налог,
сборы с которого пойдут на защиту окружающей среды,
как бы вы к этому отнеслись: положительно или отрицательно?



На основе приведенных диаграмм можно ли сделать вывод, что процент согласных с введением налога как в крупных, так и в небольших городах практически не отличается?

6. Запрет на продажу огнестрельного оружия в группе среднего возраста (30–45 лет) поддержали 77 % (из 92 опрошенных). А в старшей возрастной группе (45–60 лет) – 87,2 % (из 86 опрошенных). Подтверждают ли полученные результаты мнение о том, что поддержка запрета на продажу огнестрельного оружия не зависит от возраста? Проверьте гипотезу на уровне значимости 0,06.

7. По результатам выборочного исследования состава сотрудников компаний было сделано следующее предположение. Чем больше компания, тем она старше и тем старше её сотрудники. Для проверки

этого предположения были случайным образом выбраны компании среднего размера, т. к. большие компании чаще всего занимаются производством, и набор сотрудников осуществляется по другим критериям. Другую группу составили малые компании. Результаты исследования представлены в следующей таблице.

Возраст сотрудников, лет	18–24	25–29	30–44	45–54	55–64	65 лет и старше
«Средние» компания	9	25	45	26	38	21
«Малые» компании	14	19	21	17	12	9

Можно ли считать на уровне значимости 0,04, что малые компании имеют более молодой состав, т.е. доля молодежи в малых компаниях больше, чем в средних компаниях?

8. Если бы у Вас была возможность выбора,

Вы бы выбрали повышение в должности или прибавку к зарплате?

	Возраст, лет			
	18–24	25–34	35–44	45 и старше
повышение должности (без прибавки зарплат)	22	17	12	8
прибавку к зарплате	25	38	45	52
затрудняюсь ответить	5	6	7	5

Можно ли утверждать, что доли опрошенных, которые предпочитают повышение должности (без прибавки зарплат), в группах 18–24-летних и 25–35-летних существенно не отличаются?

4.5 Критерий χ^2

Критерий χ^2 (хи-квадрат) применяется в различных целях. Мы рассмотрим два случая:

- 1) проверка гипотезы о независимости признаков,
- 2) проверка гипотезы о виде распределения.

Критерий χ^2 основан на сравнении реальных (наблюдаемых) частот и теоретических частот, которые имели бы место, если бы проверяемая гипотеза была верна. Это различие реальных и теоретических частот может быть представлено в следующем виде:

$$\chi^2 = \sum \frac{(E - O)^2}{E}$$

где O (*observed*) – реальные частоты; E (*expected*) – теоретические частоты.

Критерий χ^2 был разработан в 1900 г. английским математиком, биологом и философом Карлом Пирсоном. Поэтому другое название критерия χ^2 – критерий Пирсона.

Проверка гипотезы о независимости признаков

С помощью критерия χ^2 мы можем определить, являются ли признаки независимыми.

При изучении взаимосвязи двух признаков или отсутствия связи данные наблюдений удобно представить в виде таблицы сопряженности признаков. Такие двумерные таблицы позволяют наглядно представить полученную информацию, систематизировать ее, сделать предварительный анализ. Таблицы сопряженности признаков также называются корреляционными, комбинационными, двумерными.

Таблица сопряженности размерности $m \times k$ имеет следующий вид.

Таблица 1

$X \backslash Y$	1 класс	2 класс		k класс	$n_{i\bullet}$
1-й класс	n_{11}	n_{12}		n_{1k}	$n_{1\bullet}$
2 –й класс	n_{21}	n_{22}		n_{2k}	$n_{2\bullet}$
.....					
m-й класс	n_{m1}	n_{m2}		n_{mk}	$n_{m\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$		$n_{\bullet k}$	n

n_{ij} – число объектов (в социологии – это, чаще всего, количество опрошенных), принадлежащих i -му классу по признаку X и, одновременно, j -му классу по признаку Y .

$n_{i\bullet} = \sum_{j=1}^k n_{ij}$ – число объектов, принадлежащих i -му классу по признаку X (без учета признака Y). Фактически это сумма значений i -ой строки.

$n_{\bullet j} = \sum_{i=1}^m n_{ij}$ – число объектов, принадлежащих j -му классу по признаку Y (без учета признака X). Фактически это итоговая сумма значений j -го столбца.

$n_{\bullet j} = \sum_{i=1}^m n_{ij}$ – число объектов, принадлежащих j -му классу по признаку Y (без учета признака X). Фактически это итоговая сумма значений j -го столбца.

Частоты $n_{i\bullet}$ и $n_{\bullet j}$ называются *маргинальными частотами*.

Справедливо равенство:

$$\sum_{i=1}^m n_{i\bullet} = \sum_{j=1}^k n_{\bullet j} = \sum_{i=1}^m \sum_{j=1}^k n_{ij} = n.$$

Поясним нашу таблицу на конкретном примере.

► **Пример.** Респондентам было предложено выразить свое отношение к следующему утверждению: «В ближайшем будущем Интернет может полностью заменить все традиционные средства массовой информации (радио, телевидение, бумажные варианты газет, журналов). Результаты опроса представлены в таблице 2.

Таблица 2

Варианты ответов	Возраст			$n_{i\bullet}$
	Молодежь	Люди среднего возраста	Люди старшего возраста	
Полностью согласен	$n_{11} = 17$	$n_{12} = 17$	$n_{13} = 5$	$n_{1\bullet} = 39$
Пожалуй, согласен	$n_{21} = 13$	$n_{22} = 24$	$n_{23} = 7$	$n_{2\bullet} = 44$
Пожалуй, не согласен	$n_{31} = 7$	$n_{32} = 21$	$n_{33} = 25$	$n_{3\bullet} = 53$
Полностью не согласен	$n_{41} = 6$	$n_{42} = 9$	$n_{43} = 18$	$n_{4\bullet} = 33$
$n_{\bullet j}$	$n_{\bullet 1} = 43$	$n_{\bullet 2} = 71$	$n_{\bullet 3} = 55$	$N = 169$

Таблица 2 имеет 4 строки ($m = 4$) и 3 столбца ($k = 3$), то есть таблица вида 4×3 . Найдем маргинальные частоты:

$$n_{1\bullet} = \sum_{j=1}^k n_{ij} = 17 + 17 + 5 = 39, \quad n_{\bullet 1} = \sum_{i=1}^m n_{ij} = 17 + 13 + 7 + 6 = 43.$$

И т.д.

$$\text{Проверка: } \sum_{i=1}^m n_{i\bullet} = \sum_{j=1}^k n_{\bullet j} = \sum_{i=1}^m \sum_{j=1}^k n_{ij} = n$$

$$\sum_{i=1}^m n_{i\bullet} = 39 + 44 + 53 + 33 = 169, \quad \sum_{j=1}^k n_{\bullet j} = 43 + 71 + 55 = 169.$$

Во 1-ой строке 3-м столбце стоит число 5 ($n_{13} = 5$). Это означает, что из всех опрошенных 5 человек являются людьми среднего возраста и они выразили полное согласие с утверждением, что «Интернет может полностью заменить все СМИ».

В последней строке в 3-м столбце стоит число 55 ($n_{33} = 55$), то есть всего в опросе приняли участие 55 человек среднего возраста.

В последнем столбце в 1-ой строке находится число 39 ($n_{1\bullet} = 39$), то есть 39 человек (разного возраста) выразили полное согласие с утверждением.

Всего было опрошено $n = 169$ человек. ■

Проверим гипотезу о независимости признаков.

Порядок применения критерия χ^2 .

1. Формулируется гипотеза $H_0 : p_{ij} = p_{i\bullet} \cdot p_{\bullet j}$ (признаки независимые), задается уровень значимости α .
2. Получают данные и оформляют их в виде таблицы сопряженности признаков.
3. Вычисляется статистика:

$$\chi^2 = n \sum_{i=1}^m \sum_{j=1}^k \frac{(n_{ij} - (n_{i\bullet} \times n_{\bullet j} / n))^2}{n_{i\bullet} \times n_{\bullet j}} = n \left(\sum_{i=1}^m \sum_{j=1}^k \frac{n_{ij}^2}{n_{i\bullet} \times n_{\bullet j}} - 1 \right)$$

При больших значениях n эта статистика имеет χ^2 – распределение с $\nu = (m - 1)(k - 1)$ степенями свободы.

4. По таблице «Процентные точки распределения χ^2 » (приложение, таблица III) находят критическое значение χ^2 – критерия при заданном уровне значимости α и числе степеней свободы ν .

Если $\chi^2 < \chi^2_{кр}(\alpha; \nu)$, то нет оснований отвергать нулевую гипотезу. Следовательно, исследуемые признаки независимы.

► **Пример (продолжение).** По результатам опроса (таблица 2) было выдвинуто предположение о том, что отношение респондента к утверждению: «В ближайшем будущем Интернет может полностью заменить все традиционные СМИ» связано с возрастом. Проверьте это предположение при уровне значимости 0,05.

Решение. Для проверки нашего предположения мы с помощью критерия χ^2 проверим гипотезу о независимости признака Y (возраст человека) и признака X (мнение человека по заданному вопросу).

1. Гипотеза $H_0: p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$

Уровень значимости $\alpha = 0,05$, критерий χ^2

2.–3 . По данным таблицы 2 находим статистику χ^2 :

$$\begin{aligned} \chi^2 &= n \left(\sum_{i=1}^m \sum_{j=1}^k \frac{n_{ij}^2}{n_{i\cdot} \times n_{\cdot j}} - 1 \right) = \\ &= 169 \left(\frac{17^2}{43 \cdot 39} + \frac{17^2}{71 \cdot 39} + \frac{5^2}{55 \cdot 39} + \right. \\ &\quad \left. + \frac{13^2}{43 \cdot 44} + \frac{24^2}{71 \cdot 44} + \frac{7^2}{55 \cdot 44} + \right. \\ &\quad \left. + \frac{7^2}{43 \cdot 53} + \frac{21^2}{71 \cdot 53} + \frac{25^2}{55 \cdot 53} + \right. \\ &\quad \left. + \frac{6^2}{43 \cdot 33} + \frac{9^2}{71 \cdot 33} + \frac{18^2}{55 \cdot 33} - 1 \right) = 29,383, \end{aligned}$$

$$\nu = (m - 1)(k - 1) = (3 - 1)(4 - 1) = 6.$$

4. По таблице «Критические точки распределения χ^2 » (таблица III приложения) находим $\chi^2_{кр} = \chi^2_{кр}(\alpha; \nu) = \chi^2_{кр}(0,05; 6) = 12,59$. Критическая область изображена на рис. 1 (с.90).

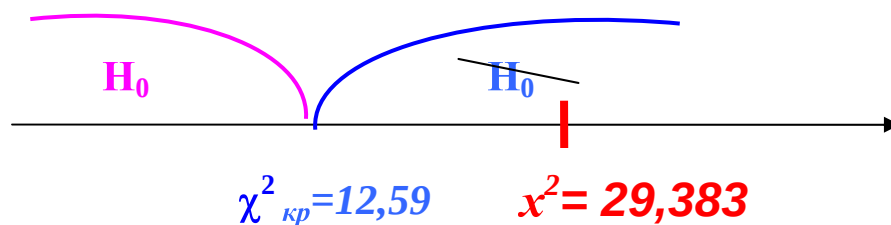


Рис. 1

Поскольку $x^2 > \chi^2_{кр}$ ($29,383 > 12,59$), то на уровне значимости 0,05 мы гипотезу H_0 (о независимости признаков) отвергаем.

Вывод: между признаками существует зависимость. То есть, на уровне значимости 0,05 подтверждается наличие связи между возрастом человека и его мнением по заданному вопросу. ■

Проверка гипотезы о равномерности генерального распределения

Критерий χ^2 используют для проверки гипотезы о виде распределения: нормального, Пуассона, экспоненциального, равномерного и т.д. Мы рассмотрим только гипотезу о равномерном распределении.

Если верна гипотеза о равномерном распределении, то все теоретические частоты $n_i^{\text{теор}}$ равны между собой:

$$n_1^{\text{теор}} = n_2^{\text{теор}} = \dots = n_k^{\text{теор}} = \frac{n}{k},$$

где $n = \sum_{i=1}^k n_i^{\text{выб}}$.

Порядок применения критерия χ^2 .

1. Формулируется гипотеза $H_0: p_1 = p_2 = p_3 = \dots = p_k$ (распределение равномерное), задается уровень значимости α .

2. Получают данные.

3. Вычисляется статистика:

$$\chi^2 = \sum_{j=1}^k \frac{(n_i^{\text{теор}} - n_i^{\text{выб}})^2}{n_i^{\text{теор}}},$$

где $n_i^{\text{выб}}$ – наблюдаемые частоты,

$$n_1^{\text{теор}} = n_2^{\text{теор}} = \dots = n_k^{\text{теор}} = \frac{n_1^{\text{выб}} + n_2^{\text{выб}} + \dots + n_k^{\text{выб}}}{k}.$$

При больших значениях n эта статистика имеет χ^2 – распределение с $\nu = k - 1$ степенями свободы.

4. По таблице «Критические точки распределения χ^2 » (приложение, таблица III) находят критическое значение χ^2 – критерия при заданном уровне значимости α и числе степеней свободы ν .

Если $\chi^2 < \chi_{\text{кр}}^2(\alpha; \nu)$, то нет оснований отвергать нулевую гипотезу. Следовательно, исследуемый признак имеет равномерное распределение.

► **Пример.** Опрос студентов группы выявил следующую ситуацию с использованием мобильных операторов.

Таблица 3

Мобильный оператор	МТС	Мегафон	Билайн	Yota
Количество студентов	18	22	19	5

Можно ли утверждать, что указанные операторы мобильной связи являются одинаково популярными среди студентов?

Решение. Для проверки нашего предположения мы с помощью критерия χ^2 проверим гипотезу о равномерном распределении.

1. Гипотеза $H_0: p_1 = p_2 = p_3 = p_4$ (распределение равномерное),

Уровень значимости $\alpha = 0,05$, критерий χ^2

2.–3 . Находим значения теоретических частот:

$$n_1^{\text{теор}} = n_2^{\text{теор}} = \dots = n_k^{\text{теор}} = \frac{18 + 22 + 19 + 5}{4} = 16$$

По данным таблицы 3 находим статистику χ^2 :

$$\begin{aligned} \chi^2 &= \sum_{j=1}^k \frac{(n_i^{\text{теор}} - n_i^{\text{выб}})^2}{n_i^{\text{теор}}} = \\ &= \frac{(16-18)^2}{16} + \frac{(16-22)^2}{16} + \frac{(16-19)^2}{16} + \frac{(16-5)^2}{16} = 10,625 \end{aligned}$$

Число степеней свободы равно $\nu = k - 1 = 4 - 1 = 3$.

4. По таблице «Критические точки распределения χ^2 » (таблица III приложения) находим $\chi_{кр}^2 = \chi_{кр}^2(\alpha; \nu) = \chi_{кр}^2(0,05; 3) = 7,815$. Критическая область изображена на рис. 1.

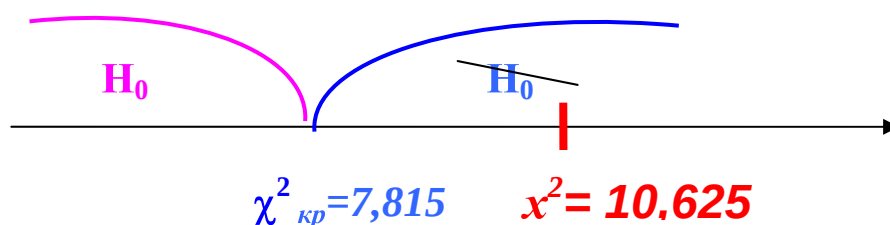


Рис. 1

Поскольку $x^2 > \chi_{кр}^2$ ($10,625 > 7,815$), то на уровне значимости 0,05 гипотеза H_0 (о равномерности распределения) отвергается.

Вывод: указанные операторы мобильной связи не являются одинаково популярными среди опрошенных студентов. ■

Контрольные вопросы.

1. Какие данные могут быть использованы в критерии χ^2 ?
2. Что такое независимость случайных величин?

3. Можно ли с помощью критерия χ^2 определить степень связи между признаками?
4. Выпишите формулу для вычисления статистики критерия сопряженности хи-квадрат.
5. Каким еще критерием (и в каком случае) можно проверить гипотезу независимости двух случайных величин?

Задачи к §4.5

1.

Любите ли Вы свою работу (ту, на которой работаете сейчас)?

Вариант ответа	Возраст			
	18–22	22–29	30–45	46–59
Очень люблю	7	21	20	29
Люблю, но без фанатизма	19	31	41	45
Отношусь спокойно	29	27	24	23
Просто ненавижу	11	6	7	5
Не хочу отвечать	5	6	7	5

- 1) Чему равно n_{22} ? Что показывает это значение?
- 2) Чему равно $n_{.2}$? Что показывает это значение?
- 3) Чему равно n_2 ? Что показывает это значение?
- 4) Сколько человек участвовало в опросе?
- 5) Сколько человек, участвующих в опросе, очень любят свою работу?
- 6) Сколько человек молодого возраста (до 30 лет) принимало участие в опросе?

7) Какова доля людей (%), которые относятся спокойно к своей работе, среди а) опрошенных в возрасте 18–22 года; б) опрошенных старше 29 лет?

8) Проверьте, подтверждают ли полученные данные связь между возрастом и любовью к работе.

2. Респондентам было предложено оценить, как изменилось их жизнь (в целом) в последний год. Результаты опроса представлены в следующей таблице.

Вариант ответа	Среднее специальное образование	Высшее и неполное высшее образование
Улучшилась 😊	16	9
Без изменения 😐	14	16
Ухудшилась ☹	8	25

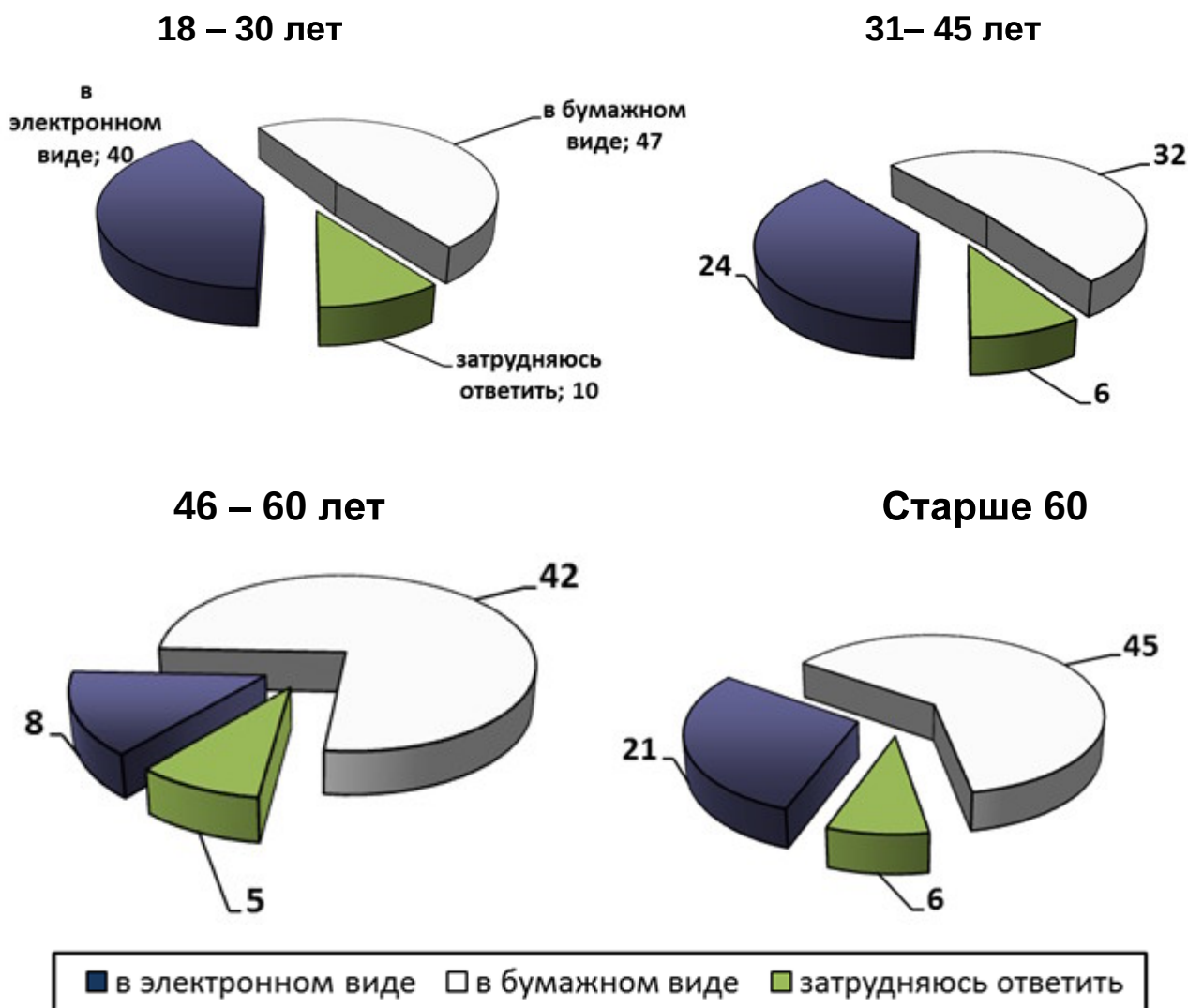
Свидетельствуют ли полученные данные о том, что отношение к жизни связано с уровнем образования? Проверьте это предположение на уровне значимости 0,05.

3. Домашняя занятость респондентов, состоящих в браке.

	Объем занятости в домашнем хозяйстве (часов в неделю)			
	Менее 10	10–19	20–39	40 и более
 Мужчины	45	30	20	22
 Женщины	8	27	42	22

Определите, действительно ли объем занятости в домашнем хозяйстве зависит от пола.

4. Вы чаще читаете книги в электронном или в бумажном виде? (человек)



Вопрос задавался тем, у кого есть электронная книга, планшет, смартфон, – отвечали 43 % респондентов.

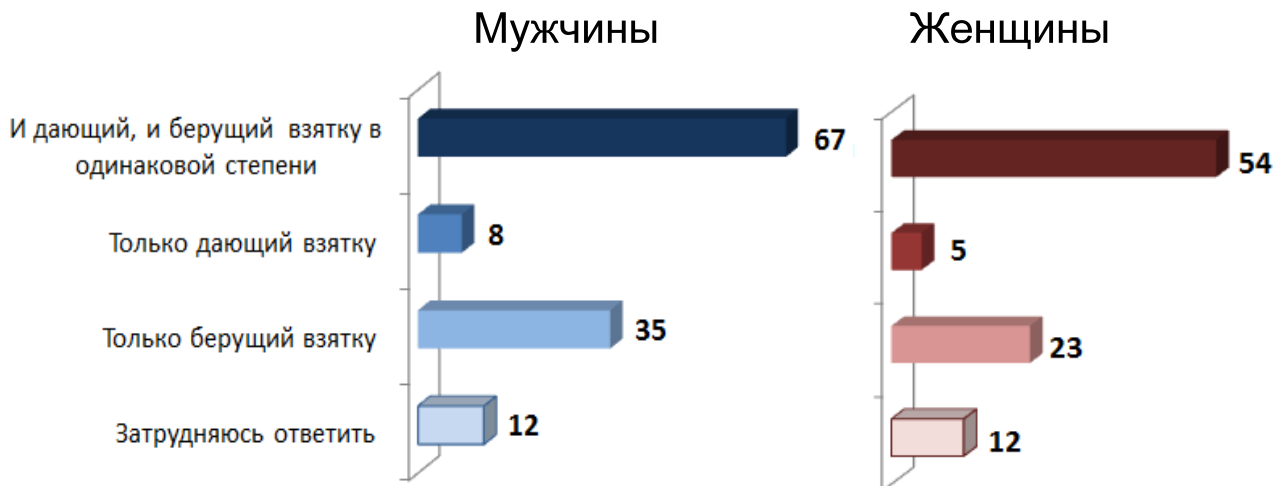
Представьте данные из диаграмм в виде таблицы сопряженности признаков. Проверьте, связаны ли с возрастом предпочтения в выборе вида книги – электронном или бумажном виде.

5. Согласно данным HR-агентств, сложнее всего найти работу людям, старше 45 лет. Для исследования была получена выборка по возрастам безработных, которые находятся в поиске работы более 3-х месяцев: 18; 27; 69; 35; 24; 31; 56; 23; 52; 19; 48; 46; 42; 64; 28; 67; 36; 49; 27; 66; 48; 55; 21; 45; 29; 30; 61; 41; 59; 44; 58; 40; 59; 51; 24; 65; 39; 61; 55; 57; 26; 68; 30; 47.

Применяя разные группировки по интервалам, проверьте гипотезу о равномерности распределения при $\alpha = 0,01$.

6. Как Вы считаете, кто виновен в процессе дачи взятки?

Ответы мужчин и женщин распределились следующим образом.



Представьте данные из диаграмм в виде таблицы сопряженности признаков. Можно ли утверждать, что пол влияет на выявление виновного в процессе дачи взятки?

7. Распределение ответов учащихся на вопрос «Занимаетесь ли Вы физической культурой и спортом?»

Варианты ответа	Юноши	Девушки
Да, регулярно	37	20
Да, время от времени, по возможности	35	36
Крайне редко	12	21
Нет, не занимаюсь	15	22

Можно ли утверждать на уровне значимости 0,01, что юноши и девушки одинаково активно занимаются спортом?

8. Менеджеру по закупкам для определения объема заказа важно знать, какие модели телефонов покупаются чаще всего. Поэтому с целью изучения покупательского спроса на телефоны было проведе-

но выборочное исследование. Результаты представлены в следующей таблице.

Тип модели	Телефон			
	А	Б	В	Г
Количество покупок	11	28	21	26

а) Можно ли утверждать, что все 4 типа модели одинаково популярны у покупателей? Уровень значимости – 0,05.

б) Можно ли утверждать, что именно модель А является значительно менее популярной относительно других моделей?

Уровень значимости – 0,05.

(Указание. Вывод можно сделать на основе результатов предыдущего пункта и проверки утверждения о том, что модели Б, В, Г одинаково популярны у покупателей.)

9. Имеются следующие данные о распределении молочных ферм по производительности труда и себестоимости молока.

Себестоимость	Производительность		
	Высокая	Средняя	Низкая
Высокая	6	13	28
Средняя	21	27	22
Низкая	35	10	5

Определите, используя критерий хи-квадрат, есть связь между производительностью труда и себестоимостью молока

10. В компании по производству молочных продуктов решили поменять упаковку для масла. Право выбора новой упаковки из 5 об-

разцов было предложено самим покупателям. Голоса распределились следующим образом.

	Образец новой упаковки				
	№1	№2	№3	№4	№5
Количество голосов	12	20	18	16	14

Какой вывод должны сделать в компании на основе полученных результатов?

11. Имеется следующая информация по предприятиям о стоимости основных производственных фондов и затратах на капитальный ремонт.

Затраты на кап. ремонт (в % к стоимости основных производственных фондов)	Среднегодовая стоимость основных производственных фондов, млн д.е.			
	5–9	9–13	13–17	17–21
1–5	5	6	31	38
5–10	7	36	18	5
10–15	13	16	6	7
15–20	21	8	7	5

Можно ли утверждать на уровне значимости 0,01, что между стоимостью основных производственных фондов и затратах на капитальный ремонт существует тесная связь?

12. Распределение заказов в фирме по месяцам первого полугодия составило: январь – 25, февраль – 38, март – 45, апрель – 32, май – 41, июнь – 43. Свидетельствуют ли эти данные о том, что фирма работает ритмично?

5. Статистическое изучение связи

Ключевые слова: функциональная зависимость, статистическая зависимость, корреляционная зависимость, факторный признак, результативный признак

В исследованиях большое место занимает определение связи между признаками. При изучении зависимости между различными явлениями и их признаками различают функциональную и статистическую зависимости.

При *функциональной связи* значение одной величины однозначно определяется значением другой и выражается математической формулой. В реальной жизни такие зависимости чаще наблюдаются в физике, биологии, химии и т.д. Например, законы Ньютона, связи между общим стажем работы и стажем работы на данном предприятии и т.д.

Но более широкое распространение в нашей жизни имеет *статистическая зависимость*. В этом случае при фиксированном значении одной величины другая имеет некоторую свободу и может принимать различные значения. Например, зависимость между стажем и производительностью труда, выпавшими осадками и полученным урожаем, ростом и весом и т.д.

Статистическая зависимость между двумя переменными, при которой каждому значению одной переменной соответствует определенное условное математическое ожидание (среднее значение) другой, называется *корреляционной*.

Понятие корреляции в принятом нами значении было введено сэром Френсисом Гальтоном в середине XIX века. Он использовал этот термин в работах по биологии, связанных с проблемами наследственности. Ф.Гальтон записывал это слово, как co-relation, то есть не просто связь (relation), а «как бы связь», соотношение.

Справедливости ради следует отметить, что этот термин использовался еще до Ф.Гальтона. В конце XVIII века французский палеонтолог Жорж Кювье применял это слово в своих работах и ввел даже термин «закон корреляции» частей и органов животных.

При анализе любого типа связи мы выделяем зависимые и независимые признаки.

Признаки, обуславливающие изменение других, связанных с ними признаков, называются *факторными (независимыми)*.

Признаки, зависимые от факторных признаков, называются *результативными (зависимыми)*.

Рассмотрим, например, зависимость между доходом и потреблением товаров. Очевидно, что независимым признаком является доход (X), а зависимым – потребление товаров (Y). То есть здесь связь «причина-следствие», $X \rightarrow Y$. Такой тип связи еще называют «направленной». В нашем примере с увеличением доходов снижается относительное потребление продовольственных товаров и увеличивается потребление непродовольственных товаров. Но следует помнить, что мы рассматриваем статистическую зависимость, и здесь не будет однозначной определенности, скорее мы определим общую тенденцию. В нашем примере возможна ситуация, когда рост доходов приведет к увеличению относительного потребления продовольственных товаров. Это может быть связано с разными привычками потребителей, неодинаковым ассортиментом продовольственных товаров, действием рекламы, новой информацией о здоровом образе жизни и т.д.

Существует другое название направленной связи. Это *асимметричная-симметричная связь*. В случае вычисления направленного коэффициента между признаками X и Y значение коэффициента для $X \rightarrow Y$ не равно, за редким исключением, значению для $Y \rightarrow X$. То есть мы не можем формально поменять X и Y местами, потому что в итоге мы получим новый коэффициент для новой связи. Также асимметрич-

ные коэффициенты обычно рассматриваются в комплексе по три: два направленных и третий, как бы их усредняющий, ненаправленный.

В случае *косвенной зависимости* признаки (X и Y) не имеют причинно-следственной связи, потому что оба являются следствием общей для них причины (признак Z , рис. 1).

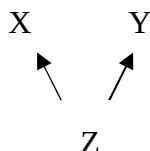


Рис. 1

Так в дореволюционной России была установлена тесная связь (корреляционная) между числом пожаров в стране и размером урожая. Число пожаров было довольно высоким в неурожайные годы. Для нас очевидно, что плохие урожаи не могут быть причиной пожаров в зданиях. В действительности же между этими явлениями существует только косвенная связь. И пожары, и урожай зависят от третьего явления – погоды. Сильная засуха приводит к плохому урожаю и способствует возникновению пожаров. Поэтому и проявляется связь между числом пожаров и урожайностью.

Следует упомянуть еще о *ложной связи*. Под ложной связью понимается чисто формальная связь между явлениями, не находящая никакого логического объяснения и основанная лишь на количественном соотношении между ними. Приведем широко известный пример ложной корреляции. Ради шутки была вычислена корреляция между числом аистов, свивших гнезда в южных районах Швеции, и рождаемостью в эти же годы в Швеции. В результате была получена положительная корреляция.

По количеству факторов зависимости бывают:

– однофакторные (парные), когда рассматривается связь двумя признаками, один из которых независимый, а другой – зависимый;

– многофакторные (множественные), когда рассматривается связь между одним зависимым признаком и несколькими независимыми.

Мы будем рассматривать только парные связи.

На сегодняшний день разработано большое число коэффициентов, которые определяют самые разнообразные виды связей практически для всех типов данных. Общее название этих коэффициентов – коэффициенты связи или меры связи. В пособии изучение корреляционной связи мы выделили в отдельный раздел.

Меры связи позволяют определять тесноту связи. Значение большинства мер находится в интервале $[-1;1]$. Некоторые меры имеют только положительные значения, поэтому интервалом их значений будет $[0;1]$. Чем ближе мера связи к ± 1 , тем сильнее связь и наоборот. Если связь отсутствует, то мера связи равна нулю.

Контрольные вопросы

1. В чем состоит отличие между функциональной и стохастической связью?
2. Какой признак называют факторным?
3. Какой признак называют результативным?
4. В каких границах находятся значения мер связи?

5.1. Корреляционный и регрессионный анализ

Ключевые слова: корреляционная зависимость, факторный признак, результативный признак, положительная корреляция, отрицательная корреляция, линия регрессии.

Корреляционная зависимость – это разновидность статистической зависимости, при которой каждому значению одной переменной

соответствует определенное условное математическое ожидание (среднее значение) другой.

Если коэффициент корреляции больше нуля, то речь идет о *положительной корреляции (прямая связь)*. Это означает, что с увеличением одного признака значение другого признака имеет тенденцию к возрастанию. Например, зависимость уровня производительности труда от стажа работы.

Отрицательной корреляции (обратной связи) соответствует коэффициент, который меньше нуля. В этом случае признаки находятся в «противофазе»: увеличение одного признака сопровождается убыванием другого. Например, если цены на потребительские товары растут, то спрос на них падает и наоборот.

Если признаки независимые, то коэффициент корреляции равен нулю. Обратное утверждение неверно. Если коэффициент равен нулю, то мы не можем сказать, что признаки независимы. Мы можем только утверждать, что между признаками нет корреляционной зависимости, но возможна другая зависимость, которую наш коэффициент не определяет.

Для описания связи между случайными признаками X и Y также используется графическое представление данных. Каждую пару чисел $(x_1, y_1), \dots, (x_n, y_n)$, где x_i — i -ое наблюдение признака X , а y_i — i -ое наблюдение признака Y , которое соответствует значению x_i можно рассматривать как координатные точки в прямоугольной системе координат. Тогда совокупность наблюдений, представленная таким образом, образует на плоскости скопление точек. Такого рода график называется *корреляционным полем* или *точечной диаграммой*. По конфигурации скопления точек мы можем сделать предварительные выводы о типе зависимости. По форме корреляционная связь бывает линейная и криволинейная. Например, если между точками можно провести прямую линию так, что точки будут достаточно близки к

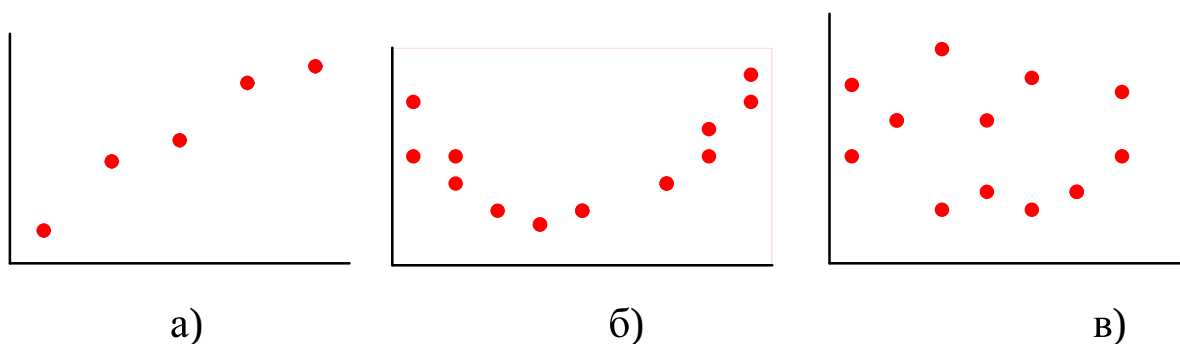


Рис. 1

ней, то речь идет о линейной зависимости (рис. 1.а). В случае рисунка 1.б зависимость наблюдается, но точно не линейная, а скорее похожая на параболическую (криволинейная зависимость). Рисунок 1.в является примером того, что между признаками точно нет зависимости. В этом случае проводить дальнейшие исследования просто нет смысла.

Основная задача *корреляционного анализа* – определение степени и направления связи между признаками.

Контрольные вопросы

1. Что представляет собой корреляционная связь?
2. При каком значении коэффициента корреляции наблюдается наиболее тесная связь?
3. При каком значении коэффициента корреляции наблюдается обратная связь между признаками?
4. При каком значении коэффициента корреляции наблюдается прямая связь между признаками?
5. Какой вывод можно сделать, если коэффициент корреляции равен нулю?
6. С какой целью используется корреляционный анализ?
7. В чем смысл коэффициента корреляции?

5.1.1. Линейный коэффициент корреляции.

Элементы регрессионного анализа

В данном разделе мы рассмотрим коэффициент корреляции, который показывает меру линейной зависимости между двумя признаками. Этот коэффициент еще называют коэффициентом корреляции Пирсона или линейным коэффициентом корреляции.

Выборочный коэффициент корреляции определяется по формуле:

$$r = \frac{\frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})}{S_x \cdot S_y}, \quad -1 \leq r \leq 1.$$

или

$$r = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \cdot \bar{y}}{S_x \cdot S_y},$$

где $\bar{x} = \frac{1}{n} \sum x_i$ – выборочное среднее случайной величины X ;

$S_x^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$ – выборочная дисперсия случайной величины X .

В случае, когда число наблюдений над признаками достаточно велико, для большей наглядности и упрощенности расчетов массив наблюдаемых значений подвергают группировке и представляют в виде корреляционной таблицы. Тогда выборочный коэффициент корреляции имеет следующий вид:

$$r = \frac{\frac{1}{n} \sum_{i=1}^k \sum_{j=1}^m n_{ij} (z_i - \bar{x})(t_j - \bar{y})}{S_x \cdot S_y},$$

где n_{ij} – частота попадания в i -й класс по признаку X и j -й класс по признаку Y , z_i и t_j – середины классов, k и m – число классов по признакам X и Y соответственно.

Если мы перейдем к условным вариантам вида

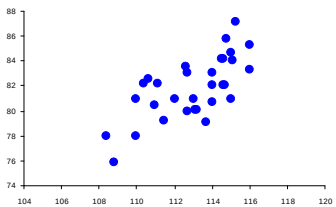
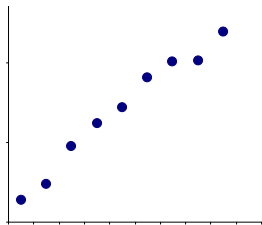
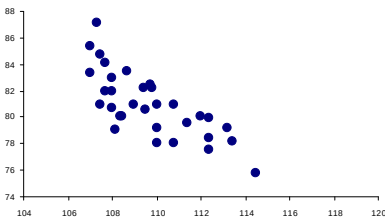
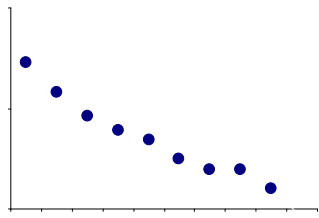
$$u_i = \frac{x_i - C_1}{d_1} \text{ и } v_i = \frac{y_i - C_2}{d_2},$$

то формула выборочного коэффициента корреляции практически не

изменится:
$$r = \frac{\frac{1}{n} \sum_i \sum_j n_{ij} u_i v_j - \bar{u} \bar{v}}{S_u \cdot S_v}.$$

В таблице 1 представлены варианты корреляционного поля и их интерпретация.

Таблица 1

 $r > 0$ Слабая прямая связь	 $r \rightarrow 1$ Строгая прямая связь
 $r < 0$ Слабая обратная связь	 $r \rightarrow -1$ Строгая обратная связь

Обычно считают, что если $|r| \geq 0,7$, то связь сильная; при $0,5 < |r| < 0,7$ – средняя связь; при $|r| < 0,5$ – слабая связь. Следует отметить, что указанная интерпретация значений коэффициента корреляции достаточно условна. Сила связи определяется сущностью анализируемых явлений. Например, в социологии значения r редко достигают границ варьирования, и связь считается сильной, если $|r| > 0,3$.

Если $|r| = 1$, то между случайными величинами X и Y существует функциональная зависимость ($Y = aX + b$). И, следовательно, по значениям одного признака мы можем предсказывать значение другого.

При $r = 0$ говорят, что с.в. X и Y некоррелированы, т.е. между ними нет корреляционной зависимости (но не исключается другой тип зависимости).

Коэффициент корреляции Пирсона является симметричным коэффициентом $r_{xy} = r_{yx}$, т.е. он только указывает на степень связи между признаками. А вот что является причиной (независимый признак), а что – следствием (зависимый признак), приходится определять исходя из предыдущих исследований или обычного здравого смысла.

▶ Пример 1. В таблице представлены данные о работниках бригады: разряд (X) и стаж работы (Y , лет). Вычислить выборочный коэффициент корреляции, построить диаграмму рассеивания.

X	1	1	2	3	6
Y	1	2	3	4	5

Решение. Чаще всего, чем больше стаж работы, тем выше квалификация рабочего и, следовательно, выше его разряд. Поэтому в нашей задаче стаж работы (X) – независимая переменная, а разряд (Y) – зависимая от X переменная.

Для построения диаграммы рассеивания на координатную плоскость нанесем точки с координатами (1; 1), (1; 2), (2; 3), (3; 4), (6; 5) (рис. 1). По конфигурации скопления точек мы можем предположить, что связь между стажем и разрядом есть, причем, скорее всего эта связь линейная и положительная.

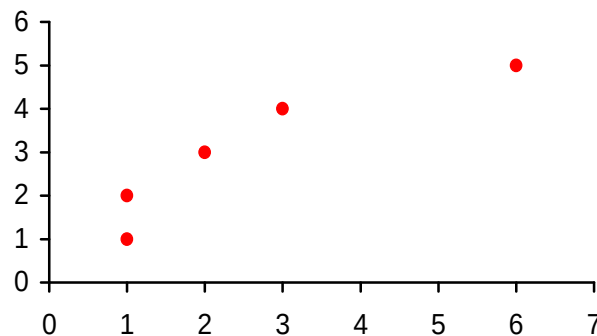


Рис. 1

Найдем выборочный коэффициент корреляции по следующей

формуле:
$$r = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \cdot \bar{y}}{S_x \cdot S_y}.$$

Для удобства представим наши вычисления в виде следующей расчетной таблицы:

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	1	1	1	1
1	2	1	4	2
2	3	4	9	6
3	4	9	16	12
6	5	36	25	30
$\Sigma = 13$	15	51	55	51

$n = 5.$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{5} \cdot 13 = 2,6; \quad S_x^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2 = \frac{1}{5} \cdot 51 - (2,6)^2 = 3,4;$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{5} \cdot 15 = 3, \quad S_y^2 = \frac{1}{n} \sum_{i=1}^n y_i^2 - (\bar{y})^2 = \frac{1}{5} \cdot 55 - 3^2 = 2,$$

$$r = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \cdot \bar{y}}{S_x \cdot S_y} = \frac{\frac{1}{5} \cdot 51 - 2,6 \cdot 3}{\sqrt{3,4 \cdot 2}} = 0,92.$$

Полученное значение коэффициента $r = 0,92$ очень близко к 1, что указывает на сильную зависимость между стажем и разрядом рабочего. ■

► **Пример 2.** Имеется следующая информация по однотипным предприятиям торговли о возрасте (продолжительности эксплуатации) типового оборудования (X , лет) и затратах (Y , тыс. д.е.) на их ремонт. Вычислить выборочный коэффициент корреляции.

Таблица 2

$X \backslash Y$	0–1,5	1,5–3,0	3,0–4,5	4,5–6,0	6,0–7,5
1–3	1	1	–	–	–
3–5	–	3	2	1	–
5–7	–	1	2	4	6
7–9	–	–	–	2	2

Решение. Признаки X и Y представлены в таблице 2 в виде интервальных рядов. Поэтому сначала мы находим середины классов

$(\tilde{x}_i, \tilde{y}_j)$. Например, середина первого интервала по признаку Y будет

$$\tilde{y}_1 = \frac{0+1,5}{2} = 0,75. \text{ В итоге получаем таблицу 3.}$$

Таблица 3

$X \backslash Y$	0,75	2,25	3,75	5,25	6,75	n_X
2	1	1	-	-	-	2
4	-	3	2	1	-	6
6	-	1	2	4	6	13
8	-	-	-	2	2	4
n_Y	1	5	4	7	8	$n = 25$

В столбце с обозначением n_X записывается в i -й строке число наблюдений только признака X , которые принадлежат i -му интервалу без учета другого признака (§4.5). Аналогично и для строки n_Y . Рассмотрим признаки X и Y по отдельности.

1) Распределение признака X :

X	2	4	6	8
n_i	2	6	13	4

В строке n_i мы записываем значения столбца n_X .

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i n_i = \frac{2 \cdot 2 + 4 \cdot 6 + 6 \cdot 13 + 8 \cdot 4}{25} = \frac{138}{25} = 5,52,$$

$$\frac{1}{n} \sum_{i=1}^n x_i^2 n_i = \frac{4 \cdot 2 + 16 \cdot 6 + 36 \cdot 13 + 64 \cdot 4}{25} = \frac{828}{25} = 33,12,$$

$$S_x^2 = \frac{1}{n} \sum_{i=1}^n n_i x_i^2 - (\bar{x})^2 = 33,12 - (5,52)^2 = 2,65, \quad S_x = \sqrt{S_x^2} = \sqrt{2,65} = 1,63.$$

2) Распределение признака Y:

Y	0,75	2,25	3,75	5,25	6,75
n_i	1	5	4	7	8

В строке n_j мы записываем значения столбца n_Y .

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i n_i = \frac{0,75 \cdot 1 + 2,25 \cdot 5 + 3,75 \cdot 4 + 5,25 \cdot 7 + 6,75 \cdot 8}{25} = \frac{117,75}{25} = 4,71,$$

$$\frac{1}{n} \sum_{i=1}^n y_i^2 n_i = \frac{0,75^2 \cdot 1 + 2,25^2 \cdot 5 + 3,75^2 \cdot 4 + 5,25^2 \cdot 7 + 6,75^2 \cdot 8}{25} = \frac{639,56}{25} = 25,58$$

$$S_y^2 = \frac{1}{n} \sum_{i=1}^n n_i y_i^2 - (\bar{y})^2 = 25,58 - (4,71)^2 = 3,4, \quad S_y = \sqrt{S_y^2} = \sqrt{3,4} = 1,84.$$

3) Вычисляем выборочный коэффициент корреляции.

$$\sum_{i=1}^4 \sum_{j=1}^5 n_{ij} x_i y_j = 1 \cdot 2 \cdot 0,75 + 1 \cdot 2 \cdot 2,25 + 3 \cdot 4 \cdot 2,25 + 2 \cdot 4 \cdot 3,75 + 1 \cdot 4 \cdot 5,25 +$$

$$+ 1 \cdot 6 \cdot 2,25 + 2 \cdot 6 \cdot 3,75 + 4 \cdot 6 \cdot 5,25 + 6 \cdot 6 \cdot 6,75 + 2 \cdot 8 \cdot 5,25 + 2 \cdot 8 \cdot 6,75 = 703,5$$

$$r_{xy} = \frac{\frac{1}{n} \sum n_{ij} x_i y_j - \bar{x} \cdot \bar{y}}{S_x S_y} = \frac{\frac{1}{25} \cdot 703,5 - 5,52 \cdot 4,71}{1,63 \cdot 1,84} = 0,71.$$

Значение выборочного коэффициента корреляции свидетельствует о средней степени зависимости между возрастом (продолжительностью эксплуатации) типового оборудования и затратами на их ремонт. ■

Основываясь только на значении выборочного коэффициента корреляции, особенно если это значение не очень близко к ± 1 , нельзя сделать вывод о достоверности корреляции между признаками. Этот вывод может быть сделан с помощью соответствующих критериев значимости корреляции. Мы рассмотрим t -критерий (Стьюдента).

Условия применения t -критерия: $(X_1, Y_1), \dots, (X_n, Y_n)$ – случайная выборка из двумерного нормального распределения со средними μ_x и μ_y , дисперсиями σ_x^2, σ_y^2 и коэффициентом корреляции ρ .

Гипотеза $H_0: \rho = 0$.

Альтернатива $H_1: \rho \neq 0$. Мы можем взять и одностороннюю альтернативу ($H_1: \rho > 0$ или $H_1: \rho < 0$), если знак корреляции может быть определен заранее.

Уровень значимости: α .

Порядок применения:

1. Принимается предположение о нормальности, формулируется гипотеза H_0 и альтернатива H_1 , задается уровень значимости α .
2. Получают двумерную выборку объема n .
3. Вычисляют статистику Стьюдента:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

где r – выборочный коэффициент корреляции; n – объем выборки.

Статистика T при справедливости нулевой гипотезы имеет t -распределение Стьюдента с $\nu = n - 2$ степенями свободы.

4. Границы критической области вычисляются в зависимости от выбранной альтернативы.

а) $H_1: \rho \neq 0$ $t_{кр} = t_{двуст\ кр}(\alpha; \nu)$. Значение $t_{кр} = t_{двуст\ кр}(\alpha; \nu)$ с ν степенями свободы находится по таблице «Критические точки распределения Стьюдента» (приложение, таблица III) по заданному уровню значимости α , помещенному в верхней строке таблицы (для двусторонней критической области). Если $|T| > t_{кр}$, то гипотеза H_0 отвергается.

б) $H_1: \rho > 0$ $t_{кр} = t_{прав\ кр}(\alpha; \nu)$. Если $T > t_{кр}$, то гипотеза H_0 отвергается.

в) $H_1: \rho < 0$ $t_{кр} = t_{лев\ кр} = -t_{пр\ кр}(\alpha; \nu)$. Если $T < t_{лев\ кр}$, то гипотеза H_0 отвергается.

Значения $t_{кр} = t_{прав\ кр}(\alpha; \nu)$, $t_{кр} = t_{лев\ кр} = -t_{пр\ кр}(\alpha; \nu)$ с ν степенями свободы находится по таблице «Критические точки распределения Стьюдента» (приложение, таблица III) по заданному уровню значимости α , помещенному в нижней строке таблицы (для односторонней критической области).

Вывод «гипотеза H_0 отвергается» означает, что коэффициент корреляции статистически значимо отличается от нуля на уровне значимости α . То есть, подтверждается наличие связи.

Пример 2 (продолжение). Проверим гипотезу о значимости корреляции, используя коэффициент корреляции из примера 2.

Решение.

1. Принимаем предположение о нормальности двумерного распределения.

Гипотеза $H_0: \rho = 0$.

Альтернатива $H_1: \rho > 0$ (т.к. по нашим предположениям здесь существует прямая связь: повышение продолжительности эксплуатации типового оборудования ведет к увеличению затрат на их ремонт).

Уровень значимости: $\alpha = 0,05$.

2.-3. В Примере 2 мы уже вычислили выборочный коэффициент корреляции: $r = 0,71$; $n = 25$. Находим статистику Стьюдента

$$t = \frac{0,71\sqrt{25-2}}{\sqrt{1-0,71^2}} = 4,84,$$

$$\nu = n - 2 = 25 - 2 = 23.$$

4. Для односторонней критической области находим по таблице «Критические точки распределения Стьюдента» значение

$t_{кр} = t_{прав ст кр}(\alpha; v) = t_{прав ст кр}(0,05; 23) = 1,71$. Получаем критическую область $(1,71; +\infty)$, которая изображена на рис.1

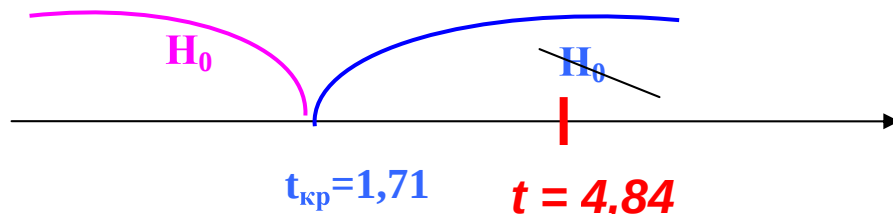


Рис. 1

Поскольку $t > t_{кр}$, т.е. значение статистики принадлежит критической области $(4,84 \in (1,71; +\infty))$, то гипотеза H_0 отвергается.

Вывод: между продолжительностью эксплуатации типового оборудования и затратах на их ремонт существует статистически значимая положительная связь ■

Как мы уже писали ранее, основной задачей корреляционного анализа является определение степени и направления связи между признаками. В регрессионном анализе также изучается связь между зависимой переменной и одной или несколькими независимыми переменными, но главным является установление формы зависимости. Поэтому основная задача *регрессионного анализа* – определение аналитического выражения связи между признаками.

Пусть переменная Y зависит от одной переменной X . При этом предполагается, что переменная X принимает заданные (фиксированные) значения, а зависимая переменная Y имеет случайный разброс из-за ошибок измерения, влияния неучтенных факторов или других причин. В этом случае уравнение взаимосвязи (парная регрессионная модель) может быть представлена в виде

$$Y = f(x) + \varepsilon, \quad (1)$$

где ε – случайная ошибка наблюдений.

Рассмотрим простейшую и наиболее часто используемую форму регрессии – парную линейную регрессию:

$$f(x) = \alpha + \beta x. \quad (2)$$

Пусть проведено n независимых наблюдений с.в. Y при значениях переменной $X = x_1, x_2, \dots, x_n$. При этом измерения с.в. Y дали следующие результаты: $y_1, y_2, \dots, y_n$. Тогда линейная регрессионная модель имеет вид:

$$y_i = \alpha + \beta x_i + \varepsilon_i. \quad (3)$$

Причем, $E(\varepsilon_i) = 0$, $D(\varepsilon_i) = \sigma^2$.

Оценкой модели (3) по выборке является уравнение регрессии:

$$\hat{y}_x = a + bx. \quad (4)$$

Принято говорить: «регрессия Y на X ».

Параметры линейного уравнения парной регрессии находят методом наименьших квадратов (МНК). Условие МНК:

$$\sum (y_i - \hat{y}_i)^2 = \sum (y_i - (a + bx_i))^2 \rightarrow \min.$$

Получаем:

$$b = b_{Y|X} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{x^2 - (\bar{x})^2} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_x^2}, \quad (5)$$

$$a = \bar{y} - b_{Y|X} \bar{x}. \quad (6)$$

Параметр $b_{Y|X}$, т.е. коэффициент при x в уравнении линейной регрессии, называется коэффициентом регрессии (коэффициентом регрессии Y на X). Коэффициент регрессии показывает, на сколько (в абсолютном выражении) изменяется в среднем значение результативного признака Y при изменении факторного признака X на единицу.

Если $b_{Y|X} > 0$, то зависимость прямая,

если $b_{Y|X} < 0$, то зависимость обратная.

До сих пор мы предполагали, что переменная Y зависит от переменной X . Рассмотрим противоположную ситуацию, когда переменная Y – независимая, а переменная X зависит от Y . Тогда уравнение регрессии X на Y :

$$\hat{x}_y = a + by,$$

$$b = b_{X|Y} = \frac{\frac{1}{n} \sum xy - \bar{x} \cdot \bar{y}}{S_y^2}, \quad a = \bar{x} - b_{X|Y} \bar{y}.$$

Между коэффициентом корреляции r и коэффициентами регрессии $b_{Y|X}$ и $b_{X|Y}$ существует взаимосвязь:

$$r = \sqrt{b_{Y|X} \cdot b_{X|Y}}.$$

Поэтому коэффициенты регрессии могут быть вычислены через коэффициент корреляции:

$$b_{Y|X} = r \frac{S_y}{S_x}, \quad b_{X|Y} = r \frac{S_x}{S_y}.$$

Проверка гипотезы о значимости коэффициента регрессии осуществляется с помощью критерия Стьюдента.

1. Гипотеза $H_0: b_{Y|X} = 0$.

Альтернатива $H_1: b_{Y|X} \neq 0$

Уровень значимости: α .

2. Получают двумерную выборку объема n .

3. Вычисляют статистику:

$$T = b_{Y|X} \frac{\sqrt{n-2}}{S_{\text{ост}}} S_x,$$

где $S_{\text{ост}}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2,$

$b_{Y|X}$ – коэффициент регрессии; n – объем выборки.

Статистика T при справедливости нулевой гипотезы имеет t -распределение Стьюдента с $\nu = n - 2$ степенями свободы.

4. Найдем границы критической области: $t_{кр} = t_{\text{двуст кр}}(\alpha; \nu).$

Значение $t_{кр} = t_{\text{двуст кр}}(\alpha; \nu)$ с ν степенями свободы находится по таблице «Критические точки распределения Стьюдента» (приложение, таблица III) по заданному уровню значимости α , помещенному в верхней строке таблицы (для двусторонней критической области).

Если $|T| > t_{кр}$, то гипотеза H_0 отвергается; коэффициент регрессии $b_{Y|X}$ статистически значим на уровне значимости α .

► **Пример (продолжение Примера 1 (§5.1.1)).** В таблице представлены данные о работниках бригады: разряд (X) и стаж работы (Y , лет). По приведенным данным а) найти уравнение регрессии Y на X ; б) проверить гипотезу о значимости коэффициента регрессии; в) построить линию регрессии Y на X ; г) оценить средний разряд рабочего при стаже 7 лет.

X	1	1	2	3	6
Y	1	2	3	4	5

Решение. Признак X (стаж работы) – независимая переменная, а признак Y (разряд) – зависимая от X переменная.

а) Предварительный анализ корреляционного поля (рис. 1, с. 104) позволяет предположить, что зависимость между стажем и разрядом является линейной. Тогда

$$\hat{y}_x = a + b_{Y|X} x.$$

Параметры линейного уравнения парной регрессии будем находить по формулам (5), (6). Для расчета параметров воспользуемся вспомогательной таблицей, уже рассмотренной нами в Примере 1 (§5.1.1). Добавим в таблицу еще одну строку – средние значения по каждому столбцу.

$n = 5$.

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	1	1	1	1
1	2	1	4	2
2	3	4	9	6
3	4	9	16	12
6	5	36	25	30
$\Sigma = 13$	15	51	55	51
$\frac{\Sigma}{n} = 2,6$	3	10,2	11	10,2

$$b = b_{Y|X} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{10,2 - 2,6 \cdot 3}{10,2 - (2,6)^2} = 0,7,$$

$$a = \bar{y} - b_{Y|X} \bar{x} = 3 - 0,7 \cdot 2,6 = 1,19.$$

Значение $b_{Y|X} = 0,7$ показывает, что с увеличением рабочего стажа на 1 год разряд в среднем возрастает на 0,7.

$$\hat{y}_x = a + b_{Y|X} x = 1,19 + 0,7x.$$

Это уравнение характеризует зависимость среднего значения разряда рабочего от его стажа.

б) Проверим гипотезу о значимости коэффициента регрессии на уровне значимости 0,05.

1. Гипотеза $H_0: b_{Y|X} = 0$.

Альтернатива $H_1: b_{Y|X} \neq 0$

Уровень значимости: $\alpha = 0,05$.

2.-3. $b_{Y|X} = 0,7; n = 5$.

Сначала найдем расчетные значения $\hat{y}_i$.

При $x_1 = 1$ вычислим значение $\hat{y}_1 = 1,19 + 0,7 \cdot 1 = 1,89$;

$x_2 = 1$ значение $\hat{y}_2 = 1,89$;

$x_3 = 2$ значение $\hat{y}_3 = 1,19 + 0,7 \cdot 2 = 2,59$;

$x_4 = 3$ значение $\hat{y}_4 = 1,19 + 0,7 \cdot 3 = 3,29$;

$x_5 = 6$ значение $\hat{y}_5 = 1,19 + 0,7 \cdot 6 = 5,39$.

$$\begin{aligned} \text{Тогда } S_{ост} &= \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} = \\ &= \sqrt{\frac{1}{5} ((1 - 1,89)^2 + (2 - 1,89)^2 + (3 - 2,59)^2 + (4 - 3,29)^2 + (5 - 5,39)^2)} = \\ &= 0,57. \end{aligned}$$

Вычислим статистику Стьюдента:

$$t = b_{Y|X} \frac{\sqrt{n-2}}{S_{ост}} S_x = \frac{0,7 \sqrt{5-2} \cdot \sqrt{3,4}}{0,57} = 3,917.$$

Число степеней свободы $\nu = n - 2 = 5 - 2 = 3$.

4. Найдем границы критической области по таблице «Критические точки распределения Стьюдента»:

$t_{кр} = t_{двуст\ кр}(\alpha; \nu) = t_{двуст\ кр}(0,05; 3) = 3,18$. Критическая область изображена на рис.1.

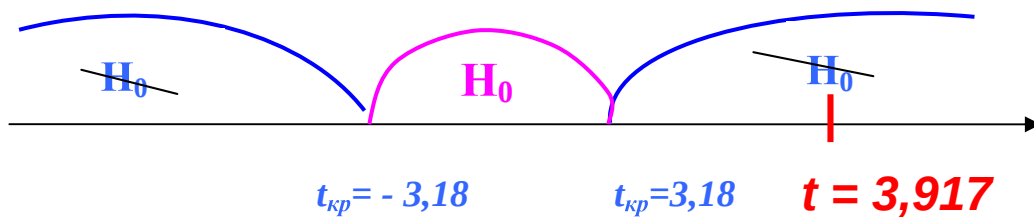


Рис. 1

Так как $|t| > t_{кр}$ ($3,917 > 3,18$), то гипотеза H_0 отклоняется. Следовательно, коэффициент регрессии $b_{Y|X} = 0,7$ является статистически значимым на уровне значимости $\alpha = 0,05$.

в) Построим линию регрессии Y на X :

$$\hat{y}_x = 1,19 + 0,7x.$$

Учитывая, что график прямой строится по любым двум точкам, выберем уже вычисленные расчетные значения $\hat{y}_i$ из (б).

Например, $(x_i; \hat{y}_i)$: $(1; 1,89)$, $(6; 5,39)$. Полученный график линии регрессии Y на X изображен на рис. 1.

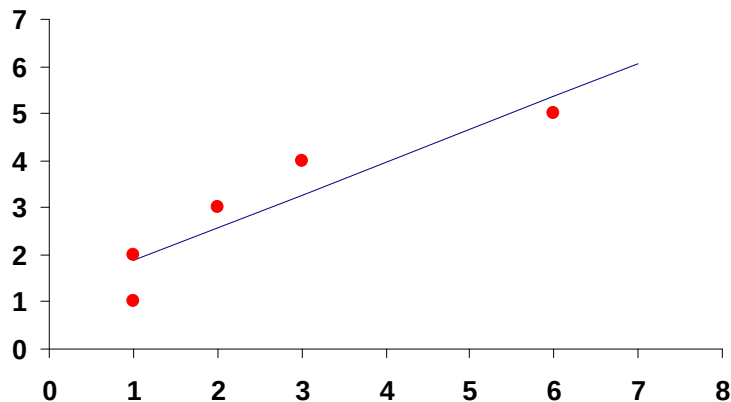


Рис. 1

г) Найдем оценку среднего разряда рабочего при стаже 7 лет. В уравнение линии регрессии Y на X подставим значение стажа $x = 7$:

$$\hat{y}_x = 1,19 + 0,7x = 1,19 + 0,7 \cdot 7 = 6,09.$$

Таким образом, мы получили, что при рабочем стаже 7 лет средний разряд рабочего будет равен 6-му разряду.

Надо отметить, что при решении подобных задач следует не забывать и обычный здравый смысл. Так, в нашей стране существуют рабочие специальности, для которых наивысший разряд равен 5. Или, например, рабочий не может сделать за смену деталей больше определенного количества (из-за возможностей оборудования и его собственных, человеческих). ■

Контрольные вопросы

1. Что показывает корреляционное поле?
2. Какими свойствами обладает коэффициент корреляции?
3. Не вычисляя, скажите, чему равен коэффициент корреляции для следующих данных: (1,4), (2,3), (3,2), (4,1).
4. Как будет располагаться линия регрессии, если коэффициент корреляции близок к 0?
5. В каком случае обе линии регрессии совпадут?
6. С какой целью проверяется гипотеза о значимости выборочного коэффициента корреляции?

Задачи к §5.1.1

В задачах 1–4 по виду корреляционного поля определите тип связи.

1.						2.				
X	2	5	6	9	11	X	9	10	12	5
Y	3	6	4	6	3	Y	6	4	7	3

3.					
X	4	3	6	7	10
Y	8	12	5	3	2

4.					
X	5	8	4	3	9
Y	9	7	9	1	10

5. В задаче представлены данные о стаже работы (X , лет) и производительностью труда (Y , деталей в час). Вычислить выборочный коэффициент корреляции, построить корреляционное поле и линию регрессии для следующей выборки:

X	8	10	5	8	9
Y	1	3	1	2	3

6. Определить, существует ли связь между возрастом самолета (X , лет) и стоимостью его годовой эксплуатации (Y , тыс. д.е.) и построить линию регрессии на основе следующих данных:

X	1	2	3	4	5
Y	4	8	12	17	22

Сделать прогноз стоимости годовой эксплуатации, если возраст самолета равен 7 лет.

7. По данным шести однотипных мясомолочных хозяйств исследуется влияние электровооруженности (X , кВт) на производительность труда (Y , тыс.руб./чел.). Статистические расчеты дали следующие результаты: $\bar{x} = 70$; $S_x = 21$; $\bar{y} = 30$; $S_y = 9$; $\sum x_i y_i = 13722.66$. Определить коэффициент корреляции между электровооруженностью и производительностью труда. Построить линию регрессии.

8. Результаты наблюдения над ростом (X) и размером обуви (Y) мужчин сведены в следующую таблицу:

X \ Y	160	165	170	175	180
38	2	1	–	–	–
39	1	3	2	–	–
40	–	1	4	1	–
41	–	–	1	3	1
42	–	–	–	1	4

По данным этой таблицы найти коэффициент корреляции, построить линию регрессии на координатной плоскости.

9. При исследовании корреляционной зависимости между стоимостью основных производственных фондов (млн д.е.) и объемом валовой продукции (млн д.е.) для 10 предприятий был получен выборочный коэффициент корреляции $r = 0,6$. Проверьте значимость этой связи при $\alpha = 0,05$.

10. Для установления статистической зависимости между себестоимостью одного изделия (Y , руб.) и величиной выпуска продукции (X , шт.) получены следующие данные: количество наблюдений $n=11$; средние значения: $\bar{y}=1,7$ р, $\bar{x} = 4$ тыс. шт.; средние квадратичные отклонения $S_x^2 = 1.2$ тыс. шт.; $S_y^2 = 0.12$; $\sum x_i y_i = 71,28$. Проверить при уровне значимости $\alpha=0,1$ существенность статистической связи (значимость коэффициента корреляции). При подтверждении значимости корреляционной связи построить уравнение регрессии.

11. Для нормирования труда проведено статистическое исследование связи между количеством изготавливаемых изделий X (шт.) и затраченным на это рабочим временем Y (мин.). Получены выборочные оценки при $n = 51$: $r = 0.8$; $S_x = 3.2$; $\bar{x} = 8$; $S_y = 8$; $\bar{y} = 40$. Можно ли считать полученную корреляционную связь достоверной при уровне значимости 0,05? Если «да», то построить уравнение регрес-

сии и вычислить предполагаемое среднее количество изделий, выпущенное за 1,5 ч.

12. Для исследования были выбраны 27 однотипных предприятий одной отрасли промышленности, которые характеризуются следующими данными.

X – балансовая прибыль (тыс. д.е.),

БАЛАНСОВАЯ прибыль – общая, суммарная прибыль предприятия, полученная за определенный период от всех видов производственной и непроизводственной деятельности предприятия, зафиксированных в его бухгалтерском балансе.

Y – объем реализованной продукции (тыс. д.е.).

По результатам наблюдений были вычислены следующие суммы:

$$\sum_{i=1}^n x_i = 190,3; \sum_{i=1}^n x_i^2 = 1519,708, \sum_{i=1}^n x_i y_i = 12835,22;$$

$$\sum_{i=1}^n y_i = 1679,9; \sum_{i=1}^n y_i^2 = 112949,7.$$

1) Найдите среднюю балансовую прибыль (тыс. д.е.), средний объем реализованной продукции (тыс. д.е.).

2) Найдите выборочный коэффициент корреляции и составьте уравнение регрессии. Проверьте значимость коэффициента регрессии.

13. Что можно сказать о взаимосвязи между признаками X и Y, если коэффициент корреляции между ними оказался равным: а) 0,15; б) 1; в) 0,95; г) 1,8; д) – 0,95?

14. По следующим данным изменения веса поросят (Y, кг) в зависимости от их возраста (X, недели) построить линию регрессии. Какой будет предположительно вес 10-недельного поросенка?

(X, Y): (0; 1,3), (1; 2,5), (2; 3,9), (3; 5,2), (4; 6,3), (5; 7,5), (6; 9,0), (7; 10,8), (8; 13,1).

5.1.2. Ранговая корреляция

Ключевые слова: ранжированная выборка, ранги, связанные (объединенные) ранги, коэффициент ранговой корреляции Спирмена, коэффициент ранговой корреляции Кендалла, коэффициент конкордации.

В исследованиях часто приходится иметь дело с признаками, которые не поддаются непосредственному измерению. Например, понятие удовлетворенности работой, местом жительства, зарплатой или оценка качества товара, его покупательского спроса и т.д. В последнее время широко используется такое понятие как «рейтинг»: рейтинг телевизионных программ, ведущих, коммерческих банков и т. п. Все эти характеристики объединяет одно: мы не можем определить их точные численные значения, но можем выразить их значения в баллах или других условных единицах. Мы можем также определить очередность значений по принципу «больше – меньше», «лучше – хуже». Т.е. признаки являются *порядковыми* или *ординальными* (§1).

Объекты (или значения признаков), расположенные в порядке возрастания или убывания их свойств образуют *ранжированную выборку*. Порядковый номер объекта в ней называется *рангом*.

При ранжировании возможна ситуация, когда разным объектам мы даем одинаковую «оценку» их признака. И получается, что одинаковым значениям приписываются разные порядковые номера и, соответственно, разные ранги. Поэтому ранг равен среднему арифметическому порядковых номеров совпадающих значений. В подобных случаях ранги называются *связанными* или *объединенными*.

Рассмотрим пример.

$n = 10$. Наша выборка: 4, 5, 8, 7, 6, 5, 7, 10, 11, 7.

Расположим данные в порядке возрастания, точнее неубывания, так как в выборке есть совпадающие значения:

x_i : 4 5 5 6 7 7 7 8 10 11

№: 1 2 3 4 5 6 7 8 9 10.

Значение $x_i = 5$ встречается два раза, поэтому ранг будет равен:

$$R = \frac{2+3}{2} = 2,5.$$

Аналогично для $x_i = 7$: $R = \frac{5+6+7}{3} = 6.$

Для остальных, несовпадающих элементов выборки, их ранги будут равны порядковым номерам.

Ранги R для данной выборки равны:

x_i : 4 5 5 6 7 7 7 8 10 11

R : 1 2,5 2,5 4 6 6 6 8 9 10. ■

Контрольные вопросы.

1. Дайте определение ранга.
2. Какие признаки называются *порядковыми*?
3. Дайте определение ранжированной выборки.
4. Как присваиваются ранги для совпадающих значений?

Задачи к §5.1.2

1. Проставьте ранги в следующей выборке: 4, 9, 14, 5, 6, 3.
2. Проранжируйте данные отдела кадров о количестве дней, пропущенных сотрудниками подразделения по больничным листам.

X : 4, 3, 2, 3, 9, 12, 3, 4, 11.

3. В таблице представлено количество баллов, набранных студентами по дисциплине «Математическая статистика» за семестр.

№ студента в списке	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Баллы за семестр	26	13	32	48	27	35	9	13	25	38	45	13	42	30

Проставьте ранги.

5.1.2.1. Коэффициент ранговой корреляции Спирмена

Коэффициент ранговой корреляции Спирмена r_s (rho) назван по имени английского психолога Чарльза Эдварда Спирмена (C.Spearman) (1904):

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \quad (1)$$

где d_i – разности рангов i -ой пары значений x_i и y_i ;

n – число сопоставляемых пар рангов или объём выборки.

$$-1 \leq r_s \leq 1.$$

Если все значения x_i и y_i имеют одинаковые ранги, то все разности рангов $d_i = 0$ и $r_s = 1$. Если ранги x_i и y_i расположены в обратном порядке, то $r_s = -1$. Таким образом, коэффициент ранговой корреляции Спирмена является мерой совпадения рангов значений x_i и y_i .

В тех случаях, когда в рядах x_i или y_i встречаются одинаковые значения, формула (1) дает несколько завышенный результат. Для более точных расчетов применяется следующая формула:

$$r_s = 1 - \frac{6 \sum d_i^2 - T_x - T_y}{\sqrt{(n^3 - n - 12T_x)(n^3 - n - 12T_y)}}, \quad (2)$$

где $T_x = \frac{1}{12} \sum (t_x^3 - t_x)$, $T_y = \frac{1}{12} \sum (t_y^3 - t_y)$;

t_x, t_y – число совпадающих значений x_i и y_i в каждой из групп, где эти значения совпадают.

Коэффициент корреляции Спирмена имеет достаточно высокую точность оценки генерального параметра ρ_s : при больших объемах выборки точность оценки с помощью коэффициента r_s составляет

91,2% по отношению к точности оценки коэффициента корреляции r . Для нормально распределенной генеральной совокупности и при достаточно большом объеме выборки ($n \geq 30$) между обоими коэффициентами существует следующее асимптотическое выражение:

$$r = 2\sin\left(\frac{\pi}{6}r_s\right).$$

Статистическая значимость коэффициента Спирмена определяется следующим образом:

а) если $n < 10$, то по таблице «Критические значения коэффициента ранговой корреляции Спирмена» (приложение, таблица V) находят $r_{s\text{ кр}}$. Если $r_s > r_{s\text{ кр}}$, то коэффициент корреляции Спирмена значим на уровне значимости α .

б) если $n \geq 10$, то можно использовать критерий Стьюдента (§5.1.1) или таблицу «Критические значения коэффициента ранговой корреляции Спирмена» (при $n \leq 40$).

 **Пример 1.** Аналитический центр получил заказ: сделать прогноз предстоящих выборов в Городскую Думу. Результаты представлены в таблице 1 (с. 129). В столбце 2 таблицы показаны оценки экспертов (по 20-бальной шкале) кандидатов в депутаты на этапе предвыборной компании. В столбце 3 представлены результаты голосования – количество поданных голосов за каждого кандидата (%). Определите, насколько прогнозы экспертов соответствует результатам выборов.

Решение. Степень «предсказания» экспертами шансов кандидатов можно определить с помощью коэффициента ранговой корреляции Спирмена.

Проранжируем значения процентных распределений для признака Y (количество голосов, поданных за кандидата; столбец 3) в порядке убывания. Самому большому значению в этом столбце – 27,8% – присваиваем ранг, равный 1. После значения 27,8 в порядке

Таблица 1

Условное имя канди- дата	Оценки экс- пертов кан- дидатов, X , балл	Количество голосов, поданных за кандидата, Y , %	Ранг канди- дата по ре- зультатам оценки экс- пертов, R_X	Ранг канди- дата по результа- там голосо- вания, R_Y	d_i	d_i^2
1	2	3	4	5	6	7
А	17	19,2	2	4	-2	4
Б	8	6,1	5	5	0	0
В	19	27,8	1	1	0	0
Г	14	23,8	3	2	1	1
Д	11	20,4	4	3	1	1
Е	4	1,7	6	6	0	0
Ж	2	1	7	7	0	0
Сумма			21	21	0	6

убывания следует значение 23,8, у него ранг будет 2. Далее значению 20,4 присваиваем ранг 3. И так далее. Найденные ранги R_Y запишем в столбец 5. Аналогично ранжируем значения признака X (столбец 2).

В столбце 6 таблицы 1 записываем разности рангов $d_i = R_{x_i} - R_{y_i}$. Если ранги рассчитаны верно, то их суммы должны быть одинаковыми, т.е. $\sum R_{x_i} = \sum R_{y_i}$ и, следовательно, $\sum d_i = 0$ (столбец 6). В нашей таблице $\sum R_{x_i} = \sum R_{y_i} = 21$

В столбце 7 возводим разности в квадрат и находим сумму $\sum d_i^2 = 8$. Подставляем найденное значение в формулу (1):

$n = 7$,

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 6}{7(7^2 - 1)} = 0,89.$$

Полученный коэффициент ранговой корреляции Спирмена показывает, что между прогнозом экспертов и результатами выборов существует достаточно тесная зависимость. Но поскольку у нас была слишком маленькая выборка (7 наблюдений), то мы не можем утверждать, что эксперты сделали достаточно хороший прогноз. Общий вывод можно сделать по результатам проверки гипотезы о равенстве коэффициента Спирмена нулю, что означает отсутствие связи.

1. Гипотеза $H_0: \rho_s = 0$.

Альтернатива $H_1: \rho_s \neq 0$

Уровень значимости: $\alpha = 0,05$.

Поскольку объем выборки $n = 7 < 10$, то для проверки гипотезы будем использовать таблицу «Критические значения коэффициента ранговой корреляции Спирмена».

2.-3. Коэффициент ранговой корреляции Спирмена $r_s = 0,89$; $n = 7$.

4. По таблице «Критические значения коэффициента ранговой корреляции Спирмена» (приложение, таблица V) находим $r_{s\text{ }kp} = 0,78$. На рис. 1 изображена критическая область.

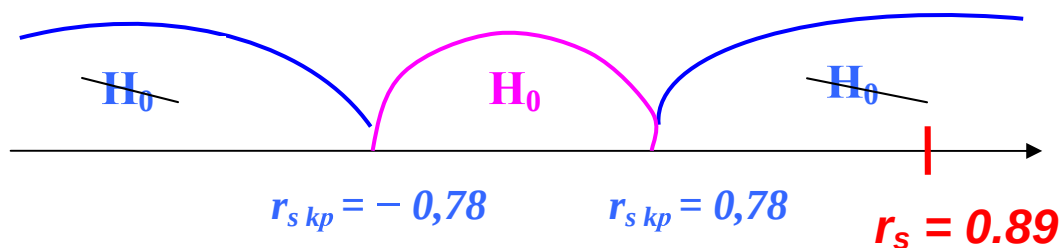


Рис. 1

Поскольку $r_s > r_{s\text{ }kp}$ ($0,89 > 0,78$), то гипотеза H_0 отклоняется.

Вывод: подтверждается достаточно тесная зависимость между прогнозами экспертов и результатами выборов. Другими словами, можно утверждать, что эксперты достаточно точно предсказали результаты выборов

5.1.2.2. Коэффициент ранговой корреляции Кендалла

Коэффициент ранговой корреляции τ (тау) М.Дж. Кендалла (M.Kendall) дает более строгую оценку корреляции рангов, чем коэффициент корреляции Спирмена. Обозначим коэффициент Кендалла через τ_k .

$$\tau_k = \frac{P - Q}{P + Q}, \quad (1)$$

где P – число рангов выше данного;

Q – число рангов ниже данного (по ряду Y).

Знаменатель этой формулы представляет собой максимально возможную сумму рангов $P + Q = \frac{1}{2}n(n-1)$. Поэтому формула (1) может быть записана в следующем виде:

$$\tau_k = \frac{P - Q}{\frac{1}{2}n(n-1)}, \quad (2)$$

где n – число объектов.

$$-1 \leq \tau_k \leq 1.$$

В случае таблицы сопряженности $m \times k$ мы имеем дело со связанными (объединенными) рангами и коэффициент τ_k Кендалла вычисляем как

$$\tau_k = \frac{2(S - D)}{\sqrt{(S + D + T_x)(S + D + T_y)}}, \quad (3)$$

где S – сумма произведений частот на сумму частот, расположенных ниже и правее данной клетки;

D – сумма произведений частот на сумму частот, расположенных ниже и левее данной клетки;

T_x – сумма произведений частот на сумму частот, расположенных в данном строке правее данной клетки.

T_y – сумма произведений частот на сумму частот, расположенных в данном столбце ниже данной клетки.

Коэффициент ранговой корреляции Кендалла показывает разность вероятностей совпадения и несовпадения порядков по обоим признакам для наугад выбранной пары объектов.

Величина коэффициента τ_k Кендалла всегда меньше, чем ρ_s . Причем при больших n ($n > 30$) и слабых или умеренных связях коэффициент связывает приближенное соотношение: $\rho_s \approx \frac{3}{2} \tau_k$.

Коэффициент корреляции Кендалла оказывается предпочтительнее с точки зрения вычислений, когда а) много рангов; б) вводятся новые объекты в процессе исследования.

Проверка гипотезы о равенстве коэффициента корреляции Кендалла нулю ($n > 10$).

1. Гипотеза $H_0: \tau_k = 0$.

Альтернатива $H_1: \tau_k \neq 0$

Уровень значимости: α .

2. Получают двумерную выборку объема n ($n > 10$).

3. Вычисляют статистику:

$$T = \tau_k \sqrt{\frac{9n(n-1)}{2(2n+5)}},$$

где τ_k – коэффициент корреляции Кендалла; n – объем выборки.

4. Найдем границы критической области: $t_{кр}$ – квантиль нормального распределения, который находится из соотношения $\Phi(t_{кр}) = 1 - \frac{\alpha}{2}$.

Если $|T| > t_{кр}$, то гипотеза H_0 отвергается; коэффициент корреляции Кендалла значим на уровне значимости α .

▶ **Пример 1 (продолжение).** Вычислить коэффициент корреляции τ_k Кендалла для данных таблицы 1(с. 129).

Решение. В примере 1 для нахождения коэффициента r_s Спирмена мы проранжировали значения таблицы 1. «Перестроим» таблицу таким образом, чтобы ранги в столбце R_x (ранг кандидата согласно мнению экспертов) располагались в порядке возрастания. А столбец R_y запишем в соответствии со столбцом R_x . Например, ранг $R_x = 1$ соответствует кандидату с условным именем «В». По результатам выборов этот кандидат сохранил свое лидирующее положение: его ранг после выборов $R_y = 1$. Следующему рангу $R_x = 2$ (кандидат с условным именем «А») соответствует $R_y = 4$. И т.д. В результате получаем следующую таблицу.

Таблица 1.1

Условное имя кандидата	Ранги рядов		P_i	Q_i
	R_x	R_y		
1	2	3	4	5
В	1	1	6	0
А	2	4	3	2
Г	3	2	4	0
Д	4	3	3	0
Б	5	5	2	0
Е	6	6	1	0
Ж	7	7	0	0
Сумма			$\sum P_i = 19$	$\sum Q_i = 2$

В дальнейших вычислениях для заполнения столбцов 4 и 5 мы будем использовать только столбец R_y .

В колонке R_y ниже ранга 1 расположены шесть рангов, превосходящих 1 (это ранги 4, 2, 3, 5, 6, 7) и ни одного ранга, который меньше 1. Поэтому $P_1 = 6$ и $Q_1 = 0$. Далее ниже ранга 4 расположены три ранга, превосходящих 4 (ранги 5, 6, 7) и два ранга, которые меньше 4 (это ранги 2, 3). Следовательно, $P_2 = 3$ и $Q_2 = 2$ и т.д. В итоге $P = P_1 + P_2 + \dots + P_6 = 6 + 3 + 4 + 3 + 2 + 1 + 0 = 19$ и $Q = 0 + 2 + 0 + 0 + 0 + 0 = 2$.

Для проверки используем известное нам равенство

$$P + Q = \frac{1}{2}n(n-1):$$

$$P + Q = 19 + 2 = 21 \text{ и } \frac{1}{2}n(n-1) = \frac{1}{2} \cdot 7(7-1) = 21.$$

По формуле (1) находим τ_K :

$$\tau_K = \frac{P - Q}{P + Q} = \frac{19 - 2}{19 + 2} = \frac{17}{21} = 0,809.$$

Мы определили, что существующую зависимость между прогнозом и результатами выборов можно охарактеризовать как связь средней степени. Как мы видим, выполняется соотношение между коэффициентами Спирмена и Кендалла: найденное значение коэффициента Кендалла меньше значения коэффициента Спирмена ($r_s = 0,89$). ■

► **Пример 2.** По данным таблицы 2 найти связь между возрастом избирателей и желанием участвовать в выборах с помощью коэффициента τ Кендалла.

Зависимость участия в выборах от возраста избирателей.

Таблица 2

Ответ	Возраст избирателей			
	18–30	31–40	41–60	Старше 60
Да	27	33	48	46
Нет	21	10	26	5

Решение. Используем формулу (3). Вычисление S , D удобнее представить в виде следующей схемы:

$$\begin{aligned}
 S = & \begin{array}{|c|c|c|c|} \hline 27 & 33 & 48 & 46 \\ \hline 21 & 10 & 26 & 5 \\ \hline \end{array} + \begin{array}{|c|c|c|c|} \hline 27 & 33 & 48 & 46 \\ \hline 21 & 10 & 26 & 5 \\ \hline \end{array} + \begin{array}{|c|c|c|c|} \hline 27 & 33 & 48 & 46 \\ \hline 21 & 10 & 26 & 5 \\ \hline \end{array} \\
 D = & \begin{array}{|c|c|c|c|} \hline 27 & 33 & 48 & 46 \\ \hline 21 & 10 & 26 & 5 \\ \hline \end{array} + \begin{array}{|c|c|c|c|} \hline 27 & 33 & 48 & 46 \\ \hline 21 & 10 & 26 & 5 \\ \hline \end{array} + \begin{array}{|c|c|c|c|} \hline 27 & 33 & 48 & 46 \\ \hline 21 & 10 & 26 & 5 \\ \hline \end{array}
 \end{aligned}$$

То есть $S = 27(10 + 26 + 5) + 33(26 + 5) + 48 \cdot 5 = 2370$,

$$D = 46(26 + 10 + 21) + 48(10 + 21) + 33 \cdot 21 = 4803$$

Величину T_x получим, если умножить частоты ячеек на сумму частот тех из них, которые стоят правее в той же строке, и сложить все $m(k-1)$ таких перекрестных произведений:

$$\begin{aligned}
 T_x = & 27(33 + 48 + 46) + 33(48 + 46) + 48 \cdot 46 + 21(10 + 26 + 5) + \\
 & + 10(26 + 5) + 26 \cdot 5 = 10040.
 \end{aligned}$$

При вычислении T_y то же самое проделывается не со строками, а со столбцами:

$$T_y = 21 \cdot 27 + 10 \cdot 33 + 26 \cdot 48 + 5 \cdot 46 = 2375.$$

Отсюда

$$\begin{aligned}
 \tau_K &= \frac{2(S - D)}{\sqrt{(S + D + T_x)(S + D + T_y)}} = \\
 &= \frac{2(4803 - 2370)}{\sqrt{(4803 + 2370 + 10040)(4803 + 2370 + 2375)}} = 0,37997
 \end{aligned}$$

Коэффициент τ_K Кендалла показывает, что существует слабая связь между возрастом избирателей и желанием участвовать в выборах. ■

Контрольные вопросы

1. Для какого типа данных предназначены коэффициенты Спирмена и Кендалла?
2. Какой из коэффициентов более строгий?
3. В чем преимущество коэффициента Кендалла?
4. В каких границах находятся значения этих коэффициентов?
5. В чем преимущество ранговых коэффициентов Спирмена и Кендалла перед коэффициентом линейной корреляции?

Задачи к §§5.1.2.1, 5.1.2.2

1. Распределение мнений респондентов по вопросу о том, какие процессы в жизни в большей степени отражают понятие «глобализация».

Варианты ответов	Респонденты	Эксперты
Передел мира в пользу сильных и богатых стран	46	49
Растущая взаимозависимость стран и народов на планете	44	59
Распространение достижений науки, техники и культуры по всему миру	36	42
Экономическая экспансия в мире транснациональных корпораций	24	45
Образование на планете единого глобального общества	22	18

Используя коэффициенты Спирмена и Кендалла, оцените степень согласованности мнений респондентов и экспертов.

2. Для определения коэффициента трудового участия работников бригады были приглашены два эксперта, которые проранжировали продукцию по качеству. Были получены следующие результаты.

Эксперт	Работник						
	1	2	3	4	5	6	7
1	1	3	6	2	4	5	7
2	4	1	5	2	6	3	7

Вычислите коэффициенты Спирмена и Кендалла. Проверьте гипотезы о значимости указанных коэффициентов при $\alpha = 0,05$.

3. Основные факторы, определившие выбор профессии среди студентов ЛАФ (Литовская академия физкультуры), КМУ (Каунасский медицинский университет), ЛСХУ (Литовский сельскохозяйственный университет), КТУ (Каунасский технологический университет).

Факторы	ЛАФ	КМУ	ЛСХУ	КТУ
Желание выразить себя в профессии	43	39	39	46
В избранной профессии вижу смысл жизни	26	35	16	36
Выбирая профессию, стремился найти контуры структуры жизни	29	29	19	38
В будущей профессии искал безопасности, удовлетворения познавательных потребностей	28	26	18	30
Возможность уехать туда, где профессия ценится выше, чем в Литве	25	25	12	8

С помощью коэффициентов Спирмена и Кендалла определите степень согласованности мнений студентов следующих вузов: а) ЛАФ и КМУ; б) ЛАФ и КТУ; в) ЛАФ и ЛСХУ.

4. На металлургическом заводе исследовалась зависимость предела прочности ($H / \text{мм}^2$) от предела текучести ($H / \text{мм}^2$). Результаты

прочности (x_i) и текучести (y_i) стали 11 марок приведены в таблице. Вычислить коэффициент линейной корреляции и коэффициент ранговой корреляции Кендалла. Проверьте гипотезы о значимости коэффициентов линейной корреляции и Кендалла.

x_i	77	96	86	92	98	53	63	80	64	66	94
y_i	81	77	76	86	53	47	36	40	49	60	80

5. Распределение ответов на вопрос «Что Вам больше всего не хватает в жизни сегодня?» (% опрошенных).

Варианты ответов	Группы работающих (средняя недельная выработка, час)			
	менее 40	40–41	42–56	свыше 56
Денег, высокой заработной платы	78	70	70	67
Жилья	9	11	11	11
Нормальной, хорошей работы по специальности	6	6	6	9
Уверенности в завтрашнем дне, стабильности	5	8	7	7
Здоровья	6	8	6	6
Свободного времени	4	4	5	9
Понимания, общения	5	6	5	4
Семьи, детей	3	4	4	3
Нормального государства, порядка, законности	2	1	2	3
Образования, знаний, опыта	2	1	1	1

С помощью коэффициентов ранговой корреляции, определите степень согласованности мнений а) в группах менее 40 и 40–41 часов; б) 42–56 и свыше 56 часов. Проверьте гипотезы о значимости коэффициентов корреляции

5.1.2.3. Множественный коэффициент корреляции W (коэффициент конкордации)

Коэффициент конкордации W показывает, насколько согласуются между собой 2 или более рядов значений переменных, представленных в порядковой шкале. В социологических исследованиях коэффициент конкордации чаще всего используется для определения степени согласованности мнений экспертов или респондентов. Если же вместо экспертов рассматривать признаки явлений, то в этом случае коэффициент конкордации можно рассматривать как показатель тесноты связи в случае множественной регрессии.

Коэффициент конкордации вычисляется по следующей формуле:

$$W = \frac{12 \cdot S}{k^2 n(n^2 - 1)}, \quad (1)$$

где k – число переменных,

n – число индивидов или категорий, которые ранжируются,

$$S = \sum_{i=1}^k (\text{сумма рангов по строке} - a)^2, \text{ а } a - \text{среднее из суммы рангов.}$$

$$0 \leq W \leq 1.$$

При наличии связанных рангов коэффициент конкордации W вычисляется по формуле:

$$W = \frac{12 \cdot S}{k^2 n(n^2 - 1) - k \sum_{j=1}^k T_j}, \quad (2)$$

$$\text{где } T_j = \sum_{j=1}^k (t_j^3 - t_j),$$

t_j – число связанных рангов по j -ой переменной.

► **Пример.** Трем респондентам было предложено указать мотивы при выборе работы в порядке предпочтения (таблица 1). Определите с помощью коэффициента конкордации, насколько согласованы мнения этих респондентов.

Таблица 1

Мотивы выбора работы	Респондент			Сумма рангов
	1-ый	2-ой	3-ий	
Хорошая зарплата	1	1	1	3
Интерес к работе	3	2,5	4	9,5
Условия труда	6	6	6	18
Трудоустройство по специальности	2	4	4	10
Коллектив	4	2,5	4	10,5
Перспективы работы	5	5	2	12
				$\Sigma=63$

Решение. Поскольку в столбцах, соответствующих мнению 2-го и 3-го респондентов, есть связанные ранги, то для нахождения коэффициента конкордации W мы будем использовать формулу (2):

$$n = 6; \quad k = 3.$$

Найдем среднее значение по сумме рангов: $a = \frac{\Sigma}{n} = \frac{63}{6} = 10,5$, тогда

$$S = (3 - 10,5)^2 + (9,5 - 10,5)^2 + (18 - 10,5)^2 + (10 - 10,5)^2 + (10,5 - 10,5)^2 + (12 - 10,5)^2 = 116.$$

Эти расчеты можно осуществить с помощью вспомогательной таблицы (таблица 2, с. 141), добавив к таблице 1 еще 2 столбца.

Таблица 2

Мотивы выбора работы	Респондент			Сумма рангов, S_i	$S_i - a$	$(S_i - a)^2$
	1-ый	2-ой	3-ий			
Хорошая зарплата	1	1	1	3	7,5	56,25
Интерес к работе	3	2,5	4	9,5	1	1
Условия труда	6	6	6	18	7,5	56,25
Трудоустройство по специальности	2	4	4	10	0,5	0,25
Коллектив	4	2,5	4	10,5	0	0
Перспективы работы	5	5	2	12	1,5	2,25
				$\Sigma=63$		$S=116$

В столбце 2-го респондента находятся два связанных ранга ($R = 2,5$), а в столбце 3-го респондента – 3 ($R = 4$). Поэтому

$$T_2 = 2^3 - 2 = 6 \text{ и } T_3 = 3^3 - 3 = 24, \quad \sum T_j = 6 + 24 = 30.$$

$$\text{Тогда } W = \frac{12 \cdot 116}{3^2 \cdot 6(6^2 - 1) - 3 \cdot 30} = 0,773, \text{ т.е. согласованность мнений респондентов достаточно велика. } \blacksquare$$

Контрольные вопросы

1. Для какого типа данных предназначен коэффициент конкордации?
2. Что определяет коэффициент конкордации?
3. В каких случаях коэффициент конкордации равен 1?
4. В каких границах находятся значения этого коэффициента?
5. Какие из перечисленных значений невозможны для коэффициента конкордации: а) $-0,4$; б) $0,4$; в) $2,4$?

Задачи к §5.1.2.3

1. Искажается ли история войны 1941–1945 гг. в некоторых публикациях печати, теле- и радиопередачах?

Варианты ответов	Возрастные группы					
	18–24	25–29	30–39	40–49	50–59	60 и старше
Да, очень часто	19	11	15	19	24	25
Иногда	20	22	23	20	18	16
Нет	2	3	3	3	2	3
Затрудняюсь ответить	8	14	8	9	6	7

Насколько совпадают мнения людей различных возрастов? Вычислите коэффициент конкордации.

2. На конкурсе вокалистов работало профессиональное жюри, а также параллельно работали зрительское жюри и жюри представителей прессы. В таблице показано распределение мест 10 конкурсантов. Проверьте, насколько совпадают мнения жюри.

Жюри	Конкурсант									
	1	2	3	4	5	6	7	8	9	10
Профессиональное	5	6	8	1	3	9	2	4	7	10
Зрительское	3	4	8	2	1	10	5	9	6	7
СМИ	4	5	8	3	1	9	6	2	7	10

3. По данным задачи №3 (§5.1.2.1 и §5.1.2.2) определите с помощью коэффициента конкордации степень согласованности мнений студентов вузов.

4. По данным задачи №5 (§5.1.2.1 и §5.1.2.2) используя коэффициент конкордации, определите, насколько согласованы мнения

людей, принадлежащих к различным группам по средней недельной выработке.

5. Определите тесноту связи между тремя факторами с помощью коэффициента конкордации.

Номер хозяйства	Посевная площадь зерновых, тыс. га	Валовый сбор, тыс. т	Внесено минеральных удобрений на 1 га посевной площади, кг
1	3,0	5,1	35
2	2,0	3,6	38
3	2,3	3,5	18
4	2,4	3,3	15
5	2,7	5,5	39
6	2,1	3,8	19
7	2,8	4,9	21
8	2,5	4,1	19
9	5,0	9,0	44
10	2,6	4,4	32
11	4,9	7,6	44
12	3,4	5,5	31

6. Семь менеджеров оценивались по методике экспертных оценок психологических характеристик личности руководителя. 15 экспертов произвели оценку каждой психологической характеристики по 5-балльной системе. В какой степени тактичность (X), требовательность (Y) и критичность (Z) взаимосвязаны?

(X,Y,Z): (70, 18, 36), (60, 17, 29), (71, 22, 40), (46, 10, 12), (58, 16, 31), (69, 19, 32), (32, 9, 13).

5.2. Ранговый бисериальный коэффициент

Ключевые слова: ранг, шкала наименований, дихотомическая шкала, ранговый бисериальный коэффициент.

Ранговый бисериальный коэффициент предназначен для изучения связи признаков, когда один признак измеряется в дихотомической шкале наименований (Y), а второй – в порядковой шкале (X).

Название коэффициента связано с тем, что мы имеем две серии данных: каждая серия для одного из значений дихотомической переменной.

Ранговый бисериальный коэффициент вычисляется по формуле:

$$r_{rb} = \frac{2}{n}(\bar{x}_1 - \bar{x}_0),$$

где n – число наблюдений,

$\bar{x}_1$ – средний ранг объектов по признаку X , имеющих значение единицы по переменной Y ,

$\bar{x}_0$ – средний ранг объектов, имеющих значение $Y = 0$.

Эта формула справедлива при отсутствии связанных рангов.

Ранговый бисериальный коэффициент изменяется в интервале от -1 до $+1$.

► **Пример.** В рамках исследования «Интернет в нашей жизни» респондентам было задано два вопроса.

1. «Вы являетесь активным пользователем социальных сетей?» (Ответ «Да» – 1, «Нет» – 0). К активным пользователем социальных сетей были отнесены респонденты, которые ежедневно используют социальные сети.

2. «Ваш возраст?»

В итоге были получены следующие данные:

(Y, X) : (1, 19), (0, 35), (1, 24), (1, 47), (0, 20), (0, 42), (1, 21), (1, 23), (0, 65), (0, 38), (0, 45).

С помощью рангового бисериального коэффициента определите, в какой степени активность в социальных сетях связана с возрастом.

Решение. Общее число опрошенных $n = 11$.

Число респондентов, ответивших «Да» ($Y = 1$, активные пользователи социальных сетей): $n_1 = 5$, а ответивших «Нет» ($Y = 0$): $n_0 = 6$.

Представим данные в виде следующей таблицы и проставим ранги.

Y	1	0	1	1	0	0	1	1	0	0	0
X	19	35	24	47	20	42	21	23	65	38	45
Ранг R_x	1	6	5	10	2	8	3	4	11	7	9

Для каждого значения признака Y выпишем соответствующие ранги по X и вычислим соответствующие ранговые средние.

Y	Ранг R_x	Ранговое среднее
$Y = 0$	6, 2, 8, 11, 7, 9	$\bar{x}_0 = \frac{6 + 2 + 8 + 11 + 7 + 9}{6} = 7,167$
$Y = 1$	1, 5, 10, 3, 4	$\bar{x}_1 = \frac{1 + 5 + 10 + 3 + 4}{5} = 4,6$

$$r_{rb} = \frac{2}{n}(\bar{x}_1 - \bar{x}_0) = \frac{2}{11}(4,6 - 7,167) = -0,467.$$

Отрицательный знак коэффициента связан с тем, что $\bar{x}_1 - \bar{x}_0 = 4,6 - 7,167 < 0$. Т.е. ранговое среднее $\bar{x}_0 = 7,167$, соответствующее значению $Y = 0$, больше рангового среднего $\bar{x}_1 = 4,6$. Это подтверждает наше предположение о том, что молодежь проявляет большую активность в социальных сетях, чем люди более старшего возраста.

Следует заметить, что мы сами определили, какому значению признака Y присваивается значение $Y = 1$, а какому – $Y = 0$. И, если бы значению $Y = 0$ соответствовали активные пользователи социальных сетей, то значение коэффициента r_{rb} было бы положительным.

Значение рангового бисериального коэффициента $|r_{rb}| = 0,467$ показывает, что связь между возрастом и активностью в социальных сетях можно охарактеризовать как среднюю. ■

Контрольные вопросы

1. Для каких шкал используется ранговый бисериальный коэффициент?
2. Что определяет ранговый бисериальный коэффициент?
3. В каком случае ранговый бисериальный коэффициент будет равен нулю?
4. Может ли ранговый бисериальный коэффициент принимать отрицательные значения?

5.3. Точечно-бисериальный коэффициент

Ключевые слова: шкала интервалов, шкала отношений, дихотомическая шкала, точечно-бисериальный коэффициент.

Точечно-бисериальный коэффициент предназначен для изучения связи признаков, когда один признак измеряется в дихотомической шкале наименований (Y), а второй – в шкале интервалов или отношений (X).

Как и в случае рангового бисериального коэффициента, мы имеем две серии данных: каждая серия для одного из значений дихотомической переменной.

Точечно-бисериальный коэффициент обозначается « r_{pb} » (от англ. *point bis series* – точечный бисериальный).

Точечно-бисериальный коэффициент вычисляется по формуле:

$$r_{pb} = \frac{\bar{x}_1 - \bar{x}_0}{nS_x} \sqrt{n_0 n_1}, \quad (1)$$

где $\bar{x}_1$ – среднее по X объектов, имеющих значение $Y = 1$;

$\bar{x}_0$ – среднее по X объектов, имеющих значение $Y = 0$;

S_x – стандартное отклонение всех n значений по X ;

n_1 – число объектов, имеющих значение $Y = 1$;

n_0 – число объектов, имеющих значение $Y = 0$;

$n = n_1 + n_0$ – число всех объектов.

Формула (1) является алгебраическим упрощением формулы коэффициента корреляции Пирсона, когда Y – дихотомическая переменная. Причем численное значение точечно-бисериального коэффициента не превышает значение коэффициента корреляции Пирсона.

Точечно-бисериальный коэффициент будет равен нулю, когда $\bar{x}_1 = \bar{x}_0$. Другими словами, когда средние по обеим сериям наблюдений равны.

► **Пример** (продолжение **Примера §5.2**). Используя точечно-бисериальный коэффициент, определите, действительно ли активность в социальных сетях связана с возрастом.

Решение. Исходные данные следующие.

Y	1	0	1	1	0	0	1	1	0	0	0
X	19	35	24	47	20	42	21	23	65	38	45

Число респондентов, ответивших «Да» ($Y = 1$, активные пользователи социальных сетей): $n_1 = 5$, а ответивших «Нет» ($Y = 0$): $n_0 = 6$.

Для каждого значения признака Y выпишем соответствующие значения X (возраст) и вычислим соответствующие средние.

Y	X	Среднее
$Y = 0$	35, 20, 42, 65, 38, 45	$\bar{x}_0 = \frac{35 + 20 + 42 + 65 + 38 + 45}{6} = 40,83$
$Y = 1$	19, 24, 47, 21, 23	$\bar{x}_1 = \frac{19 + 24 + 47 + 21 + 23}{5} = 26,8$

Средний возраст людей, причисляющих себя к активным пользователям социальных сетей, равен 26,8 лет, а для второй группы – 40,83 лет.

Средний возраст всех опрошенных равен: $\bar{x} = \frac{19 + 35 + \dots + 45}{11} = 34,46$.

$$S_x^2 = \frac{1}{n} \sum x_i^2 - \bar{x}^2 = \frac{1}{11} (19^2 + 35^2 + \dots + 45^2) - (34,46)^2 = 194,61.$$

$$\text{Тогда } r_{pb} = \frac{\bar{x}_1 - \bar{x}_0}{n S_x} \sqrt{n_0 n_1} = \frac{26,8 - 40,83}{11 \sqrt{194,61}} \sqrt{6 \cdot 5} = -0,5.$$

Мы получили отрицательный точечно-бисериальный коэффициент $r_{pb} = -0,5 < 0$ потому, что $\bar{x}_1 < \bar{x}_0$. Т.е. средний возраст активных пользователей социальных сетей оказался меньше, чем средний возраст людей, которые не причисляют себя к активным пользователям социальных сетей.

Поскольку $r_{pb} = -0,5$, то можно говорить о средней степени связи между возрастом и активностью в социальных сетях. ■

Контрольные вопросы

1. Для каких шкал используется точечно-бисериальный коэффициент?
2. Что определяет точечно-бисериальный коэффициент?

3. Чем отличаются точечно-бисериальный и рангово-бисериальный коэффициенты?
4. В каком случае точечно-бисериальный коэффициент будет равен нулю?

Задачи к §§5.2, 5.3

1. Руководство компании с целью повышения трудовой дисциплины провело замеры утренних опозданий сотрудников.

Время опозданий на работу, мин.	10	11	5	4	12	15	3	6	8	7	30
Пол	М	Ж	М	Ж	М	Ж	М	М	Ж	Ж	Ж

Определите с помощью точечно-бисериального и рангового бисериальный коэффициентов, зависит ли от пола сотрудника время опозданий на работу? Другими словами, действительно ли женщины опаздывают на значительно большее время.

2. Как известно, работа социолога – это наполовину математика, а наполовину – философия. Можно ли утверждать, что в исследуемой студенческой группе это мнение «принято на вооружение»? Студентам было задано два вопроса: «Собираетесь ли Вы работать социологом?» (Ответ «Да» – 1, «Нет» – 0) и «Сколько Вы получили баллов за тест по математике?» ($max = 20$ баллов). Получены следующие данные.

(X,Y): (1,16), (0,8), (0,12), (1,14), (0,19), (1,9), (1,17), (0,15), (1,20), (0,13), (0,18), (0,7).

Вычислите точечно-бисериальный и ранговый бисериальный коэффициенты.

3. Проверьте справедливость следующего предположения: «Люди, работающие с удовольствием, делают более качественную продукцию».

X – «Нравится ли Вам Ваша работа?» («Да» – 1, «Нет» – 0),

Y – коэффициент качества, который определяет контролер (%).

(X,Y): (1,71), (0,52), (0,41), (1,81), (0,65), (0,53), (1,63), (1,92).
(1,72), (0,48), (1,42), (0,52).

4. В период предвыборной компании был проведен анализ настроений электората. В результате были получены следующие данные.

X – «Собираетесь ли Вы участвовать в выборах?» («Да» – 1, «Нет» – 0);

Y – возраст (лет).

(X,Y): (1,54), (1,48), (0,18), (0,36), (0,60), (0,45), (1,65), (1,58), (1,38),
(1,55), (1,40), (0,37), (0,32), (1,34), (0,33), (1,49), (1,46), (1,18), (0,50), (0,64).

Используя точечно-бисериальный и ранговый бисериальный коэффициенты, определите, связано ли намерение участвовать в выборах с возрастом человека.

5. С помощью точечно-бисериального и рангового бисериального коэффициентов определите, есть ли связь между рабочим стажем в данной организации и тем, насколько работнику нравится его работа.

X – рабочий стаж (лет),

Y – ответ на вопрос «Нравится ли Вам Ваша работа?»

Стаж	5	16	10	8	9	11	15	4	7	6	12
Ответ	да	да	нет	да	да	нет	да	да	да	нет	нет

В ходе исследования опрашивались работники, имеющие стаж не менее 4 лет.

6. Рассчитайте точечно-бисериальный и ранговый бисериальный коэффициенты для **Примера**, разобранный в §§5.2.1, 5.2.2, исключив значение, соответствующее возрасту 65 лет. Как при этом изменятся коэффициенты?

7. Обучение персонала является второй по распространенности статьей расходов HR-бюджетов после расходов на подбор кадров. Даже в условиях непростого финансового положения компаний корпоративное обучение остается востребованным. Обычно главным критерием эффективности является рост производительности, другой важный показатель – приток новых клиентов.

В компании решили проверить, какая из программ повышения квалификации является более эффективной – Программа А или Программа Б. Для участия в эксперименте были выбраны сотрудники, имеющие одинаковые интегральные показатели эффективности. Часть сотрудников компании проходила переподготовку по Программе А, а другая часть – по Программе Б. По результатам года эксперт проранжировал эффективность работы сотрудников, которые прошли повышение квалификации.

Код имени сотрудника	В.Е.	С.А.	В.Б.	Г.Ю.	А.Д.	П.О.	М.Е.	Е.А.	П.П.	Н.У.	С.Л.
Программа повышения квалификации	А	В	А	А	В	А	В	В	А	В	В
Ранг	1	7	4	2	3	5	8	10	9	6	11

Можно ли утверждать, что обучение по Программе А в большей степени способствовало повышению эффективности работы?

5.4. Коэффициенты связи, основанные на критерии χ^2

Ключевые слова: мера средней квадратической сопряженностью, коэффициент взаимной сопряженности Пирсона, коэффициент P' , коэффициент Чупрова, коэффициент Крамера.

С помощью критерия χ^2 мы устанавливаем только факт зависимости между признаками, но не можем определить, насколько сильно связаны между собой эти признаки. Для определения интенсивности связи нужно использовать соответствующие меры связи, некоторые из которых мы рассмотрим в этом разделе.

Статистика X^2 (§4.5) является мерой различия между наблюдаемыми и теоретическими частотами, которая приходится на все n объекты наблюдения.

1. Мера различия, приходящаяся на одно наблюдение, называется *средней квадратической сопряженностью*:

$$\phi^2 = \frac{X^2}{n}.$$

$$0 \leq \phi^2 < +\infty.$$

Недостатком этого коэффициента является отсутствие верхней границы.

2. *Коэффициент взаимной сопряженности Пирсона*:

$$P = \sqrt{\frac{X^2}{X^2 + n}} = \sqrt{\frac{\phi^2}{\phi^2 + 1}}.$$

$$0 \leq P < 1.$$

Коэффициент P не достигает 1. Чтобы устранить этот недостаток, были предложены два варианта: формулы 3 и 4.

$$3. \quad P' = \frac{P}{C_{\max}},$$

где
$$C_{\max} = \sqrt{\frac{\min(m-1, k-1)}{1 + \min(m-1, k-1)}}.$$

Если таблица квадратная ($m = k$), то
$$C_{\max} = \sqrt{\frac{m-1}{m}}.$$

Недостаток коэффициента P' : зависимость от числа категорий.

4. *Коэффициент А.А. Чупрова:*

$$T = \sqrt{\frac{X^2}{n\sqrt{(m-1)(k-1)}}}.$$

$$0 \leq T \leq 1.$$

Коэффициент T представляет отношение средней фактической сопряженности признаков к ее максимально возможной величине.

Коэффициент Чупрова достигает максимального значения +1 в случае полной связи только для квадратной таблицы ($m = k$).

Обозначение коэффициента «Т» связано с начальной буквой фамилии А.А. Чупрова (1824 – 1926) – А.А. Tchuprov.

5. *Коэффициент Крамера.*

В отличие от коэффициента Чупрова коэффициент Г.Крамера может достигать значения +1 для любой таблицы:

$$K = \sqrt{\frac{X^2}{n \cdot \min(m-1, k-1)}}.$$

Значение коэффициента K не зависит от размерности таблиц. Очевидно, для квадратных таблиц коэффициенты Крамера и Чупрова совпадают, а в остальных случаях $K > T$.

Статистическая значимость приведенных выше коэффициентов определяется по значимости X^2 (см. §4.5): если значим X^2 , то значимы и производные коэффициенты.

► **Пример.** Найти для Таблицы 2 (§4.5) меру средней квадратической сопряженности, коэффициент взаимной сопряженности Пирсона, коэффициент P' , коэффициент Чупрова, коэффициент Крамера.

Решение. В примере (§4.5) мы нашли для Таблицы (4×3) :

$$x^2 = 29,383; \quad n = 169.$$

$$1. \quad \varphi^2 = \frac{X^2}{n} = \frac{29,383}{169} = 0,17;$$

$$2. \quad P = \sqrt{\frac{X^2}{X^2 + n}} = \sqrt{\frac{29,383}{29,383 + 169}} = 0,38.$$

$$3. \quad \text{Так как } m = 4, k = 3, \text{ то } C_{\max} = \sqrt{\frac{\min(4-1, 3-1)}{1 + \min(4-1, 3-1)}} = \sqrt{\frac{2}{1+2}} = 0,816;$$

$$P' = \frac{P}{C_{\max}} = \frac{0,38}{0,816} = 0,471;$$

$$4. \quad T = \sqrt{\frac{X^2}{n \sqrt{(m-1)(k-1)}}} = \sqrt{\frac{29,383}{169 \cdot \sqrt{(4-1)(3-1)}}} = 0,27;$$

$$5. \quad K = \sqrt{\frac{X^2}{n \cdot \min(m-1, k-1)}} = \sqrt{\frac{29,383}{169 \cdot 2}} = 0,29.$$

Таким образом, связь между возрастом человека и его мнением по заданному ему вопросу («Вы согласны с тем, что в ближайшем будущем Интернет может полностью заменить все традиционные СМИ?») характеризуется как слабая. ■

Контрольные вопросы.

1. В чем недостаток коэффициента Чупрова?
2. В каких случаях коэффициенты Чупрова и Крамера совпадают?
3. В чем недостаток коэффициента взаимной сопряженности Пирсона?
4. Зависит ли значение коэффициента Чупрова от размерности таблиц?

Задачи к §5.4

1. Что бы Вы предпочли, если бы могли выбирать?

	Возраст, лет				
	18–24	25–34	35–44	45–59	60 и старше
Небольшой, но твёрдый заработок и уверенность в завтрашнем дне	27	38	42	52	53
Много работать и хорошо зарабатывать, пусть даже без особых гарантий на будущее	39	46	42	32	18
Иметь собственное дело, вести его на свой страх и риск	27	14	9	6	5
Затрудняюсь ответить	5	4	5	4	24

1) Найдите меру средней квадратической сопряженностью, коэффициенты: взаимной сопряженности Пирсона, P' , Чупрова, Крамера.

2) Найдите все 5 коэффициентов, исключив из исходной таблицы самую молодую группу опрошенных – 18–24 г.

2. Для таблицы сопряженности признаков, состоящей из 4 строк и 3 столбцов, была проверена гипотеза об отсутствии связи между признаками на уровне значимости 0,05. Статистика критерия равна $\chi^2 = 18,225$, всего наблюдений $n = 300$. Если подтвердится зависимость признаков, то определите меру средней квадратической сопряженности, коэффициент взаимной сопряженности Пирсона, коэффициент P' , коэффициент Чупрова, коэффициент Крамера.

3. Найти для задач §4.5, в которых подтверждена зависимость признаков, меру средней квадратической сопряженности, коэффициенты: взаимной сопряженности Пирсона, P' , Чупрова, Крамера.

5.5. Коэффициенты ассоциации и контингенции

Ключевые слова: четырехклеточная таблица сопряженности, коэффициент ассоциации, коэффициент контингенции.

В этом разделе мы рассмотрим только простейший случай: связь между двумя альтернативными признаками. Мерой альтернативных признаков является наличие или отсутствие их у объектов исследования. Например, человек может иметь высшее образование или не иметь высшего образования, изучать английский язык или не изучать его, сдать зачет по математической статистике или не сдать зачет и т.д.

При исследовании сопряженности двух альтернативных признаков исходные экспериментальные данные представляют в виде четырехклеточной (четырёхпольной) таблицы сопряженности. По своему содержанию таблица 2×2 аналогична таблице, которая была введена в §4.5. Для простоты изложения внутриклеточные частоты мы обозначим через a , b , c , d и получим следующую таблицу.

Таблица 1

Признак Y \ Признак X	Наличие (+)	Отсутствие (–)	Всего
Наличие (+)	a	b	$a+b$
Отсутствие (–)	c	d	$c+d$
Всего	$a+c$	$b+d$	n

Рассмотрим конкретный пример.

▶ **Пример.** Руководство спортивного канала инициировало опрос с целью определить основной состав своей телеаудитории. По предыдущим исследованиям было известно, что спортивные передачи смот-

рят в основном мужчины. Поэтому респондентам был задан вопрос: «За последний месяц Вы смотрели хотя бы 1 спортивную передачу?»

Исследуем, в какой степени связаны пол респондента и просмотр спортивных передач с помощью таблицы 2.

Таблица 2

Пол \ Ответ на вопрос	Да	Нет	Всего
	Да	Нет	Всего
Мужчины	$a = 45$	$b = 19$	64
Женщины	$c = 28$	$d = 46$	74
Всего	73	65	138

Из таблицы 2 мы видим, что всего опрашивалось 138 человек, из них 64 мужчины и 74 женщины. Из опрошенных смотрели хотя бы одну спортивную передачу за последний месяц 73 человека: 45 мужчин и 28 женщин. ■

Для таблицы 2×2 создано много коэффициентов, определяющих зависимость между признаками. Мы рассмотрим только два из них: коэффициенты ассоциации и контингенции.

1. Коэффициент ассоциации (или коэффициент Юла (G. Yule))

$$Q = \frac{ad - bc}{ad + bc}.$$

$$-1 \leq Q \leq 1.$$

Свойства коэффициента ассоциации Q:

а) Коэффициент равен единице либо когда $c = 0$ (схема 1.а, с. 158), либо $b = 0$ (схема 1.б, с. 158).

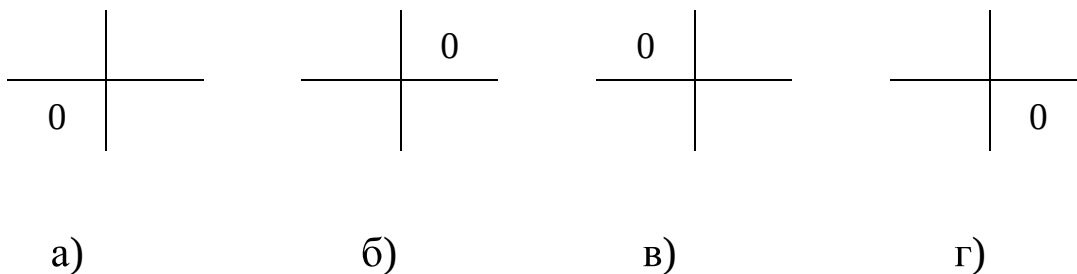


Схема 1

Представим, что $b = 0$ в таблице 2. Это означает, что среди опрошенных не было мужчин, которые бы не посмотрели хотя бы одной спортивной телепередачи за последний месяц. Другими словами, если телезритель – мужчина, то он обязательно посмотрел хотя бы одну спортивную передачу. Обратное утверждение неверно: если телезритель – женщина, то она могла посмотреть или не посмотреть спортивную передачу. Таким образом, Q – показатель односторонней связи.

б) $Q = -1$, если $a = 0$ (схема 1.в) или $d = 0$ (схема 1.г).

Наши выводы будут противоположными выводам из а).

$a = 0$: если телезритель – мужчина, то он за последний месяц не посмотрел даже одной спортивной телепередачи;

$d = 0$: если телезритель – женщина, то она обязательно за последний месяц посмотрела хотя бы одну спортивную телепередачу.

2. Коэффициент контингенции

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}.$$

$$-1 \leq \Phi \leq 1.$$

Свойства коэффициента контингенции Φ :

а) Коэффициент контингенции равен 1, если $b = c = 0$.

б) $|\Phi| \leq |Q|$

в) Если $|\Phi| \geq 0,5$, то считают, что надежно установлена двусторонняя связь.

г) Для таблицы 2×2 коэффициент Φ связан с χ^2 следующим соотношением

$$\Phi = \sqrt{\frac{\chi^2}{n}}.$$

 **Пример.** Для данных, представленных в таблице 2, найти коэффициенты ассоциации и контингенции.

Решение.

1) Вычислим коэффициент ассоциации:

$$Q = \frac{ad - bc}{ad + bc} = \frac{45 \cdot 46 - 19 \cdot 28}{45 \cdot 46 + 19 \cdot 28} = 0,59.$$

Таким образом, мы установили, что связь есть. Причем эта связь односторонняя: пол телезрителя влияет на его телевизионные предпочтения.

2) Найдем коэффициент контингенции Φ :

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} = \frac{45 \cdot 46 - 19 \cdot 28}{\sqrt{64 \cdot 73 \cdot 65 \cdot 74}} = 0,324.$$

Так как $\Phi = 0,324 < 0,5$, то, согласно свойству (в), в нашем примере нет двусторонней связи. Таким образом, в нашем примере мы можем считать надежно установленной только одностороннюю связь ($Q = 0,59$). ■

Контрольные вопросы

1. При каких значениях коэффициента контингенции считают надежно установленной двухстороннюю связь?

2. Как коэффициент контингенции связан со статистикой критерия сопряженности признаков?

3. Коэффициент ассоциации является показателем односторонней или двухсторонней связи?

4. Коэффициенты ассоциации и контингенции являются коэффициентами направленной или ненаправленной связи?

5. Для каких таблиц вычисляются коэффициенты ассоциации и контингенции?

Задачи к §5.5

1. Кого считать молодым? Росстат определяет возраст молодежи до 30 лет. А что об этом думают сами россияне? Полученные результаты опроса представлены в следующей таблице. В категории «возраст» выделены 2 группы: «младше 22 лет» и «старше 22 лет». Это связано с тем, что наибольшие изменения в восприятии возраста наблюдаются чаще всего в возрасте 22 года, когда человек получает образование, приобретает профессиональные навыки.

Вариант ответа	Возраст респондента	
	Младше 22 лет	Старше 22 лет
Да	15	89
Нет	68	25

Вычислите коэффициенты ассоциации и контингенции. Дайте оценку полученным результатам.

2. Опрос на тему «Самым эффективным способом решения проблемы воздействия автомобилизации (пробки на дорогах) является развитие системы общественного транспорта» дал следующие результаты. Из 120 владельцев автомобилей, участвующих в опросе, 45 человек согласны с утверждением. В то время как среди граждан, не имеющих личный автотранспорт, это утверждение поддержали 111

человек. Всего в опросе приняло участие 276 человек. Определите, влияет ли на мнение респондента наличие автомобиля.

3. Домашняя работа всегда считалась уделом женщин. Но в современном мире традиционные роли мужчины и женщины трансформируются, наблюдается размывание их различий. В большей степени эти изменения наблюдаются в городской среде. А как сейчас воспринимаются распределение домашних обязанностей в семье? Респондентам было предложено выразить свое согласие/несогласие со следующим утверждением: «Муж и жена должны выполнять домашнюю работу поровну». Были получены следующие результаты.

Мужчины		Женщины	
Да	Нет	Да	Нет
16	45	32	36

Оцените с помощью коэффициентов ассоциации и контингенции степень влияния пола респондента на его отношение к домашним обязанностям.

4. «Работа – это необходимость. Была бы возможность, вообще бы не работал». Ответы распределились следующим образом.

Уровень зарплаты	Вариант ответа	
	Да	Нет
Зарплата ниже среднего	52	31
Зарплата выше среднего	26	45

Связано ли желание работать с уровнем заработной платы?

5. В рамках изучения жизненных стратегий россиян респондентам было предложено определить свою позицию к следующему утверждению: «В жизни очень многое зависит от самого человека». Пред-

полагалось, в частности, что с возрастом категоричность суждений и абсолютная вера в собственные силы имеет тенденцию к снижению. Результаты опроса сведены в следующую таблицу.

Вариант ответа	Возраст			
	18–22	22–29	30–45	46–59
Полностью согласен	42	22	16	12
Пожалуй, согласен	27	37	24	18
Пожалуй, не согласен	13	16	25	18
Полностью не согласен	7	11	17	20

1) С помощью критерия независимости признаков χ^2 проверьте наличие связи между возрастом и отношением человека к своим возможностям.

2) Преобразуйте исходную таблицу в таблицу 2×2 за счет объединения столбцов и строк групп. С помощью критерия независимости признаков χ^2 проверьте наличие связи между возрастом и отношением человека к своим возможностям. Оцените с помощью коэффициентов ассоциации и контингенции степень связи между возрастом и мнением по данному вопросу.

(Предлагается в категории «Возраст» выделить две группы: первая группа «Молодые» (возраст 18–29 лет) и 2-я группа «Люди среднего и старшего возраста» (возраст 30–59 лет). В категории «Вариант ответа»: «Согласен» и «Не согласен».)

6. Как известно, мотивация играет очень большую роль в действиях человека. В студенческой группе было проведено исследование о выявлении связи между учебной мотивацией и успеваемостью по математике. Для определения уровня развития мотивации был использован тест-опросник выявления направленности и уровня развития внут-

ренной мотивации учебной деятельности студентов. В этом тесте введена 20-балльная шкала, на основе которой определяется внешняя мотивация (0–10 баллов) и внутренняя мотивация (11–20 баллов). Успеваемость определялась по результатам контрольной работы (КР), максимальная оценка за которую составляла 10 баллов. Интерпретация оценок была следующая: 0–6 баллов – недопустимый уровень, 7–10 допустимый уровень.

№	Баллы КР	Уровень мотивации	№	Баллы КР	Уровень мотивации	№	Баллы КР	Уровень мотивации
1	7	9	11	7	19	21	4	15
2	5	4	12	8	7	22	3	2
3	6	12	13	6	8	23	9	6
4	9	15	14	10	7	24	9	15
5	9	9	15	6	14	25	1	7
6	5	8	16	8	18	26	3	12
7	6	7	17	2	9	27	5	18
8	3	6	18	4	4	28	10	14
9	8	13	19	4	18			
10	8	17	20	10	15			

Составьте четырехклеточную таблицу (2×2) и вычислите помощью коэффициенты ассоциации и контингенции. Подтверждается ли предположение о том, что между мотивацией к изучению математики и успеваемостью по математике существует сильная взаимосвязь?

7. По результатам опроса на тему «Что Вы думаете по поводу обязательного тестирования школьников на наркотики?» число родителей, поддерживающих обязательное тестирование школьников на наркотики, составило 56 человек. А количество родителей, не поддер-

живающих данную инициативу, оказалось 25 человек. Среди респондентов, не имеющих детей, «За» высказались 38 человек, а 22 – «Против». Влияет ли статус «родитель» на мнение по этому вопросу?

8. Из 63 молодых сотрудников организации (возраст до 30 лет), участвовавших в опросе, третья часть опрошенных заявили, что в целом удовлетворены условиями работы. В то время как среди сотрудников среднего и старшего возраста довольными условиями работы оказались более половины (45 человек). Общее число опрошенных составило 149 человек. Определите с помощью коэффициентов ассоциации и контингенции, влияет ли возраст на оценку условий работы.

9. Опрос на тему «Как Вы собираетесь провести летний отпуск?» выявил следующие тенденции. Часть опрошенных предполагают куда-нибудь поехать: пляжный отдых, активный отдых, санаторий, экскурсии. Другая часть опрошенных – отдыхать дома, на даче. Основной причиной проведения летнего отпуска дома респонденты называют нехватку средств. Другими причинами являются преклонный возраст и состояние здоровья. Небольшая часть респондентов в качестве причины называет наличие маленьких детей. Определите, в какой степени связаны доходы респондента и организация летнего отдыха.

Место отдыха Уровень доходов	Дома, на даче	Поездка	Всего
Ниже среднего	27	15	42
Выше среднего	27	35	62
Всего	54	50	104

ПРИЛОЖЕНИЕ

Таблица I

Нормальное распределение $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$

x	Сотые доли									
	0	1	2	3	4	5	6	7	8	9
0,0	0,5	0,504	0,508	0,512	0,516	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,591	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,648	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,67	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,695	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,719	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,758	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,791	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,834	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,877	0,879	0,881	0,883
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,898	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,937	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,975	0,9756	0,9761	0,9767

x	Сотые доли									
	0	1	2	3	4	5	6	7	8	9
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,983	0,9834	0,9838	0,9842	0,9846	0,985	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,989
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,992	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,994	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,996	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,997	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,998	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,999	0,999
3,1	0,999	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

Таблица II

Критические точки распределения χ^2

$\alpha \backslash \nu$	0,99	0,95	0,90	0,10	0,05	0,02	0,01	0,001
1	0,0 ³ 157	0,00393	0,0158	2,706	3,841	5,412	6,635	10,827
2	0,0201	0,103	0,211	4,605	5,991	7,824	9,210	13,815
3	0,115	0,352	0,584	6,251	7,815	9,837	11,345	16,266
4	0,297	0,711	1,064	7,779	9,488	11,688	13,277	18,467
5	0,544	1,145	1,610	9,236	11,070	13,388	15,086	20,515
6	0,872	1,635	2,204	10,645	12,592	15,033	16,812	22,457
7	1,239	2,167	2,823	12,017	14,067	16,622	18,475	24,322
8	1,646	2,733	3,490	13,362	15,507	18,168	20,090	26,125
9	2,088	3,325	4,168	14,684	16,919	19,679	21,666	27,877
10	2,558	3,940	4,865	15,987	18,307	21,161	23,209	29,588
11	3,053	4,575	5,578	17,275	19,675	22,618	24,725	31,264
12	3,571	5,226	6,304	18,549	21,026	24,054	26,217	32,909
13	4,107	5,892	7,042	19,812	22,362	25,472	27,688	34,528
14	4,660	6,571	7,790	21,064	23,685	26,873	29,141	36,123
15	5,229	7,261	8,547	22,307	24,996	28,259	30,578	37,697
16	5,812	7,962	9,312	23,542	26,296	29,633	32,000	39,252
17	6,408	8,672	10,085	24,769	27,587	30,995	33,409	40,790
18	7,015	9,390	10,865	25,989	28,869	32,346	34,805	42,312
19	7,633	10,117	11,651	27,204	30,144	33,687	36,191	43,820
20	8,260	10,851	12,443	28,412	31,410	35,020	37,566	45,315
21	8,897	11,591	13,240	29,615	32,671	36,343	38,932	46,797
22	9,542	12,338	14,041	30,813	33,924	37,659	40,289	48,268

Окончание таблицы II

$\nu \backslash \alpha$	0,99	0,95	0,90	0,10	0,05	0,02	0,01	0,001
23	10,196	13,091	14,848	32,007	35,172	38,968	41,638	49,728
24	10,856	13,848	15,659	33,196	36,415	40,270	42,980	51,179
25	11,524	14,611	16,473	34,382	37,652	41,566	44,314	52,62(1
26	12,198	15,379	17,292	35,563	38,885	42,856	45,642	54,052
27	16,151	18,114	20,703	40,113	44,140	46,963	46,963	55,476
28	13,565	16,928	18,939	37,916	41,337	45,419	48,278	56,893
29	14,256	17,708	19,768	39,087	42,557	46,693	49,588	58,302
30	14,953	18,493	20,599	40,256	43,773	47,962	50,892	59,703
36	23,269	25,643	28,735	47,212	50,999	55,489	58,619	67,985
40	26,509	29,051	32,345	55,759	60,436	63,691	63,691	73,402
46	26,657	31,439	34,215	58,641	62,830	67,771	71,201	81,400
50	29,707	34,764	37,689	63,167	67,505	72,613	76,154	86,661
56	34,350	39,801	42,937	69,919	74,468	79,815	83,513	94,461
60	37,485	43,188	46,459	74,397	79,082	84,580	88,379	99,607
66	42,240	48,305	51,770	81,085	85,965	91,681	95,626	107,258
70	45,442	51,739	55,329	85,527	90,531	96,388	100,425	112,317

Критические точки распределения Стьюдента

Число степеней свободы v	Уровень значимости α (двусторонняя критическая область)					
	0, 10	0,05	0,02	0,01	0,002	0,001
1	6,31	12,7	31,82	63,7	318,3	637,0
2	2,92	4,30	6,97	9,92	22,33	31,6
3	2,35	3,18	4,54	5,84	10,22	12,9
4	2,13	2,78	3,75	4,60	7,17	8,61
5	2,01	2,57	3,37	4,03	5,89	6,86
6	1,94	2,45	3,14	3,71	5,21	5,96
7	1,89	2,36	3,00	3,50	4,79	5,40
8	1,86	2,31	2,90	3,36	4,50	5,04
9	1,83	2,26	2,82	3,25	4,30	4,78
10	1,81	2,23	2,76	3,17	4,14	4,59
11	1,80	2,20	2,72	3,11	4,03	4,44
12	1,78	2,18	2,68	3,05	3,93	4,32
13	1,77	2,16	2,65	3,01	3,85	4,22
14	1,76	2,14	2,62	2,98	3,79	4,14
15	1,75	2,13	2,60	2,95	3,73	4,07
16	1,75	2,12	2,58	2,92	3,69	4,01
17	1,74	2,11	2,57	2,90	3,65	3,96
18	1,73	2,10	2,55	2,88	3,61	3,92
19	1,73	2,09	2,54	2,86	3,58	3,88
20	1,73	2,09	2,53	2,85	3,55	3,85
21	1,72	2,08	2,52	2,83	3,53	3,82
22	1,72	2,07	2,51	2,82	3,51	3,79
23	1,71	2,07	2,50	2,81	3,49	3,77
24	1,71	2,06	2,49	2,80	3,47	3,74
25	1,71	2,06	2,49	2,79	3,45	3,72
26	1,71	2,06	2,48	2,78	3,44	3,71
27	1,71	2,05	2,47	2,77	3,42	3,69
28	1,70	2,05	2,46	2,76	3,40	3,66
29	1,70	2,05	2,46	2,76	3,40	3,66
30	1,70	2,04	2,46	2,75	3,39	3,65
40	1,68	2,02	2,42	2,70	3,31	3,55
60	1,67	2,00	2,39	2,66	3,23	3,46
120	1,66	1,98	2,36	2,62	3,17	3,37
∞	1,64	1,96	2,33	2,58	3,09	3,29
	0,05	0,025	0,01	0,005	0,001	0,0005
	Уровень значимости α (односторонняя критическая область)					

Процентные точки распределения Фишера-Снедекора
(F-распределения)

Таблица IV.1

F-распределение: верхние 5%-ные точки

$v_1 \backslash v_2$	1	2	3	4	5	6	7	8	9	10
1	161,4	199,5	215	224,6	230,2	234,0	236,8	238,9	240,5	241,9
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	285
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32

Продолжение таблицы IV.1

$v_1 \backslash v_2$	1	2	3	4	5	6	7	8	9	10
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83

$v_1 \backslash v_2$	12	15	20	24	30	40	60	120	∞
1	243,9	245,9	248,0	249,1	250,1	251,2	252,2	253,3	254,3
2	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93

Окончание таблицы IV.1

$v_1 \backslash v_2$	12	15	20	24	30	40	60	120	∞
9	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	2,12	2,04	1,96	1,91	1,87	1,87	1,82	1,77	1,65
29	2,01	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Таблица IV.2

F-распределение: верхние 1%-ные точки

$v_1 \backslash v_2$	1	2	3	4	5	6	7	8	9	10
1	4052	4999,5	5403	5625	5764	5859	1928	5981	6022	6056
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05
6	13,73	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13

Продолжение таблицы IV.2

$v_1 \backslash v_2$	1	2	3	4	5	6	7	8	9	10
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47
∞	6,63	4,61	3,78	3,32	3,0	2,80	2,64	2,51	2,41	2,32

$v_1 \backslash v_2$	12	15	20	24	30	40	60	120	∞
1	6106	6157	6209	6235	6261	6287	6313	6339	6336
2	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	90,50
3	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	14,37	14,20	14,12	13,43	13,84	13,75	13,65	13,56	13,46
5	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00

Окончание таблицы IV.2

$v_1 \backslash v_2$	12	15	20	24	30	40	60	120	∞
15	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	2,96	2,81	2,66	2,58	2,50	2,42	2,31	2,21	2,13
27	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,01
30	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,18
∞	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Таблица IV.3

F-распределение: верхние 0,1%-ные точки

$v_1 \backslash v_2$	1	2	3	4	5	6	7	8	9	10
1	4053*	5000*	5404*	5625*	5764*	5859*	5981*	6023*	6056*	6107*
2	998,5	999,0	999,2	999,2	999,3	999,3	999,4	999,4	999,4	999,4
3	167,0	148,5	141,1	137,1	134,6	132,8	131,6	130,6	129,9	129,2
4	74,14	61,25	56,18	53,44	51,71	50,53	49,66	49,00	48,47	48,05
5	47,18	37,12	33,20	31,09	29,75	28,84	28,16	27,64	27,24	26,92
6	35,51	27,00	23,70	21,92	20,81	20,03	19,46	19,03	18,69	18,41
7	29,25	21,69	18,77	17,19	16,21	15,52	15,02	14,63	14,33	14,08
8	25,42	18,49	15,83	14,39	13,44	12,86	12,40	12,04	11,77	11,54
9	22,86	16,39	13,90	12,56	11,71	11,13	10,70	10,37	10,11	9,89
10	21,04	14,91	12,55	11,28	10,48	9,92	9,52	9,20	8,96	8,75
11	19,69	13,81	11,56	10,35	9,58	9,05	8,66	8,35	8,12	7,92
12	18,64	12,97	10,80	9,63	8,89	8,38	8,00	7,71	7,48	7,29
13	17,81	12,31	10,21	9,07	8,35	7,86	7,49	7,21	6,98	6,80
14	17,14	11,78	9,73	8,62	7,92	7,43	7,08	6,80	6,58	6,40
15	16,59	11,34	9,34	8,25	7,57	7,09	6,74	6,47	6,26	6,08
16	16,12	10,97	9,00	7,94	7,27	6,81	6,46	6,19	5,98	5,81
17	15,72	10,66	8,73	7,68	7,02	6,56	6,22	5,96	5,75	5,58
18	15,38	10,39	8,49	7,46	6,81	6,35	6,02	5,76	5,56	5,39
19	15,08	10,16	8,28	7,26	6,62	6,18	5,85	5,59	5,39	5,22
20	14,82	9,95	8,10	7,10	6,46	6,02	5,69	5,44	5,24	5,08
21	14,59	9,77	7,94	6,95	6,32	5,88	5,53	5,31	5,11	4,95
22	14,38	9,61	7,80	6,81	6,19	5,76	5,44	5,19	4,99	4,83
23	14,19	9,47	7,67	6,69	6,08	5,65	5,33	5,09	4,89	4,73
24	14,03	9,34	7,55	6,59	5,98	5,55	5,23	4,99	4,80	4,64

* Эти значения надо умножить на 100,

Продолжение таблицы IV.3

$v_1 \backslash v_2$	1	2	3	4	5	6	7	8	9	10
25	13,88	9,22	7,45	6,49	5,88	5,46	5,15	4,91	4,71	4,56
26	13,74	9,12	7,36	6,41	5,80	5,38	5,07	4,83	4,64	4,48
27	13,61	9,02	7,27	6,33	5,73	5,31	5,00	4,76	4,57	4,41
28	13,50	8,93	7,19	6,25	5,66	5,24	4,93	4,69	4,50	4,35
29	13,39	8,85	7,12	6,19	5,59	5,18	4,87	4,64	4,45	4,29
30	13,29	8,77	7,05	6,12	5,53	5,12	4,82	4,58	4,39	4,24
40	12,61	8,25	6,60	5,70	5,13	4,73	4,44	4,21	4,02	3,87
60	11,97	7,76	6,17	5,31	4,76	4,37	4,09	3,87	3,69	3,54
120	11,38	7,32	5,79	4,95	4,42	4,04	3,77	3,55	3,38	3,24
∞	10,83	6,91	5,42	4,62	4,10	3,74	3,47	3,27	3,10	2,96

$v_1 \backslash v_2$	12	15	20	24	30	40	60	120	∞
1	6107*	6158*	6209*	6235*	6261*	6287*	6313*	6340*	6366*
2	999,4	999,4	999,4	999,5	999,5	999,5	999,5	999,5	999,5
3	128,3	127,4	126,4	125,9	125,4	125,0	124,5	124,0	123,5
4	47,41	46,76	46,10	45,77	45,43	45,09	44,75	44,40	44,05
5	26,42	25,91	25,39	25,14	24,87	24,60	24,33	24,06	23,79
6	17,99	17,56	17,12	16,89	16,67	16,44	16,21	15,99	15,75
7	13,71	13,32	12,93	12,73	12,53	12,33	12,12	11,91	11,70
8	11,19	10,84	10,48	10,30	10,11	9,92	9,73	9,53	9,33
9	9,57	9,24	8,90	8,72	8,55	8,37	8,19	8,00	7,81
10	8,45	8,13	7,80	7,64	7,47	7,30	7,12	6,94	6,76
11	7,63	7,32	7,01	6,85	6,68	6,52	6,35	6,17	6,00
12	7,00	6,71	6,40	6,25	6,09	5,93	5,76	5,59	5,42
13	6,52	6,23	5,93	5,78	5,63	5,47	5,30	5,14	4,97

Окончание таблицы IV.3

$v_1 \backslash v_2$	12	15	20	24	30	40	60	120	∞
14	6,13	5,85	5,56	5,41	5,25	5,10	4,94	4,77	4,60
15	5,81	5,54	5,25	5,10	4,95	4,80	4,64	4,47	4,31
16	5,55	5,27	4,99	4,85	4,70	4,54	4,39	4,23	4,06
17	5,32	5,05	4,78	4,63	4,48	4,33	4,18	4,02	3,85
18	5,13	4,87	4,59	4,45	4,30	4,15	4,00	3,84	3,67
19	4,97	4,70	4,43	4,29	4,14	3,99	3,84	3,68	3,51
20	4,82	4,56	4,29	4,15	4,00	3,86	3,70	3,54	3,38
21	4,70	4,44	4,17	4,03	3,88	3,74	3,58	3,42	3,26
22	4,58	4,33	4,06	3,92	3,78	3,63	3,48	3,32	3,15
23	4,48	4,23	3,96	3,82	3,68	3,53	3,38	3,22	3,05
24	4,39	4,14	3,87	3,74	3,59	3,45	3,29	3,14	2,97
25	4,31	4,06	3,79	3,66	3,52	3,37	3,22	3,06	2,89
26	4,24	3,99	3,72	3,59	3,44	3,30	3,15	2,99	2,82
27	4,17	3,92	3,66	3,52	3,38	3,23	3,08	2,92	2,75
28	4,11	3,86	3,60	3,46	3,32	3,18	3,02	2,86	2,69
29	4,05	3,80	3,54	3,41	3,27	3,12	2,97	2,81	2,64
30	4,00	3,75	3,49	3,36	3,22	3,07	2,92	2,76	2,59
40	3,64	3,40	3,15	3,01	2,87	2,73	2,57	2,41	2,23
60	3,31	3,08	2,83	2,69	2,55	2,41	2,25	2,08	1,89
120	3,02	2,78	2,53	2,40	2,26	2,11	1,95	1,76	1,54
∞	2,74	2,51	2,27	2,13	1,99	1,84	1,66	1,45	1,00

Критические значения коэффициента
ранговой корреляции Спирмена

n	α		n	α		n	α	
	0,05	0,01		0,05	0,01		0,05	0,01
5	0.94	—	17	0,48	0.62	29	0.37	0.48
6	0.85	—	18	0,47	0.60	30	0,36	0,47
7	0.78	0.94	19	0,46	0.58	31	0.36	0.46
8	0.72	0,88	20	0,45	0,57	32	0.36	0.45
9	6.68	0.83	21	0,44	0.56	33	0.34	0.45
10	0.64	0,79	22	0,43	0,54	34	0.34	0.44
11	0.61	0.76	23	0.42	0.53	35	0.33	0.43
12	0.58	0.73	24	0.41	0.52	36	0.33	0.43
13	0,56	0,70	25	0.49	0,51	37	0.33	0.43
14	0.54	0.68	26	0,39	0.50	38	0.32	0.41
15	0.52	0.66	27	0,38	0,49	39	0.32	0.41
16	0.50	0.64	28	0.38	0.48	40	0.31	0.40

ГЛОССАРИЙ

Альтернативная (конкурирующая гипотеза) – это гипотеза H_1 , которая противоречит нулевой гипотезе.

Вариационный ряд – последовательность вариантов, записанных в возрастающем порядке.

Выборка – это совокупность случайно отобранных объектов.

Выборочное среднее $\bar{X}$ – это среднее арифметическое всех вариантов выборки.

Выборочная дисперсия S^2 – это средний квадрат отклонения значений вариант от их средней арифметической.

Выборочное стандартное отклонение (или выборочное среднее квадратическое отклонение) – это корень квадратный из выборочной дисперсии

Гистограмма – фигура, состоящая из прямоугольников, высота которых пропорциональна числу наблюдений, попавших в данные интервалы числовой прямой.

Дискретный вариационный ряд – это перечень вариант и соответствующих им частот или относительных частот.

Доверительный интервал для параметра θ – это интервал $(\hat{\theta}_1, \hat{\theta}_2)$ со случайными концами, накрывающий истинное значение θ с вероятностью не меньшей $1 - \alpha$, т.е. $P(|\theta - \hat{\theta}| < \Delta) \geq 1 - \alpha$.

Доверительный уровень (доверительная вероятность, надежность) – это вероятность γ того, что доверительный интервал накроет истинное значение параметра.

Интервальная оценка – это оценка, которая определяется двумя числами – концами интервалов.

Интервальный вариационный ряд – это соответствие между интервалами и частотами, которые равны сумме частот вариантов, относящихся к интервалу.

Критерий K – это правило, по которому определяется, принять или отклонить гипотезу H_0 .

Критическая область – это множество всех значений статистики критерия, при которых нулевая гипотеза отклоняется

Математическая статистика – раздел математики, изучающий математические методы сбора, систематизации, обработки и интерпретации результатов наблюдений с целью выявления статистических закономерностей.

Медиана ($\hat{Me}$) – это значение признака, которое приходится на центральный член вариационного ряда.

Мода ($\hat{Mo}$) – это значение варианты, частота или относительная частота которой имеет наибольшее значение.

Мощность критерия – вероятность $(1 - \beta)$ не допустить ошибку 2-го рода, т.е. отвергнуть гипотезу H_0 , когда она неверна.

Надежность – см. доверительный уровень.

Область принятия гипотезы – это множество всех значений статистики критерия, при которых нулевая гипотеза принимается

Объем выборки – это число элементов выборки.

Оценка (статистическая оценка) – это функция от наблюдаемых случайных величин, принимающая значения в пространстве возможных значений оцениваемого параметра.

Ошибка второго рода – это принятие ложной гипотезы H_0 ,

Ошибка первого рода – это отклонение гипотезы H_0 , когда она верна.

Полигон – это ломаная кривая, соединяющая точки $(x_1, h_1), (x_2, h_2), \dots, (x_r, h_r)$, абсцисса которых соответствует варианту или середине ин-

тервала, а ордината пропорциональна частоте варианты или соответствующего интервала.

Предельная ошибка выборки Δ – это наибольшее отклонение оценки $\hat{\theta}$ от оцениваемого параметра θ , которое возможно с заданным доверительным уровнем γ .

Признаки зависимые – см. *признаки результативные*.

Признаки независимые – см. *признаки факторные*.

Признаки результативные (зависимые) – это признаки, зависимые от факторных признаков.

Признаки факторные (независимые) – это признаки, которые обуславливают изменение других, связанных с ними признаков.

Ранг – это порядковый номер объекта в ранжированной выборке.

Ранжированная выборка – это выборка, в которой объекты (или значения признаков) расположены в порядке возрастания или убывания их свойств.

Связь корреляционная – статистическая связь между двумя переменными, при которой каждому значению одной переменной соответствует определенное условное математическое ожидание (среднее значение) другой.

Связь функциональная – связь, при которой значение одной величины однозначно определяется значением другой и выражается математической формулой.

Статистика – это любая функция от наблюдений.

Статистической гипотезой, обозначаемой H , называется любое предположение относительно вида или параметра распределения с.в. X .

Частота – это число наблюдений варианты.

Частость (относительная частота) – это отношение частоты к объему выборки.

ИНФОРМАЦИОННЫЕ ИСТОЧНИКИ

Основная литература

1. Ватутин В.А., Ивченко Г.И., Медведев Ю.И. и др. Теория вероятностей и математическая статистика в задачах: учебное пособие для вузов. – 2-е изд., испр. – М.: Дрофа, 2003. – 328 с.: ил.
2. Горелова Г.В., Кацко И.А. Теория вероятностей и математическая статистика в примерах и задачах с применением Excel: учебное пособие для вузов. – Изд. 3-е, доп. и перераб. – Ростов-на-Дону: Феникс, 2005. – 480 с.
3. Кремер Н.Ш. Теория вероятностей и математическая статистика. – 3-е изд., перераб. и доп. – М.: ЮНИТИ-ДАНА, 2009. – 551 с.
4. Общий курс высшей математики для экономистов: Учебник под ред. В.И. Ермакова. – М.:ИНФРА-М, 2003. – 575 с.

Дополнительная литература

5. Айвазян С.А., Мешалкин Л.Д., Енюков И.С. Прикладная статистика. Основы моделирования и первичная обработка данных. – М.: Финансы и статистика, 1985. – 472 с.
6. Аптон Г. Анализ таблиц сопряженности. – М.: Финансы и статистика, 1982 (Upton G.J.G. The analysis of cross-tabulated data. N.-Y.: J.Wiley&Sons, 1978). – 143 с.: ил.
7. Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. – М.: Прогресс, 1976. – 493 с.
8. Гмурман В.Е. Теория вероятностей и математическая статистика. — 9-е изд., стер. — М.: Высш. шк., 2003. — 479 с., ил.
9. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. – М.: Высшая школа, 1998. – 400 с., ил.

10. Елисеева И.И. Статистические методы измерения связей. – Л.: ЛГУ, 1982. – 136 с.
11. Калинина В.Н., Панкин В.Н. Математическая статистика. – М.: Высшая школа. – 4-е изд., испр. — М., 2002. — 336 с., ил.
12. Каштанова Е.К. Меры связи, применяемые в социологических исследованиях. – Казань. Изд-во КГУ, 2006. – 72 с.
13. Каштанова Е.К. Сборник задач по математической статистике для студентов экономического факультета. – Казань. Изд-во КГУ, 2006. – 84 с.
14. Кимбл Г. Как правильно пользоваться статистикой / пер. с англ. Б.И.Клименко. – Финансы и статистика, 1982. – 294 с., с ил.
15. Паниотто В.И., Максименко В.С. Количественные методы в социологических исследованиях. – Киев: Наукова Думка: 1982. – 272 с.
16. Справочник по прикладной статистике / под ред. Э. Ллойда и У. Ледермана. Том 2. – М.: Финансы и статистика, 1990. – 526 с.
17. Татарова Г.Г. Методология анализа данных в социологии. – М., 1998. – 223 с.
18. Толстова Ю.Н., Математико-статистические модели в социологии (математическая статистика для социологов): учебное пособие / Гос. ун-т – Высшая школа экономики. – 2-е изд. – М.: Изд. дом ГУ ВШЭ, 2008. – 243 с.
19. Толстова Ю.Н., Куликова А.А., Рыжова А.В. Математическая статистика для социологов. Задачник. Гос. ун-т – Высшая школа экономики. – М.: Изд. дом ГУ ВШЭ, 2010. – 183 с.
20. Толстова Ю.Н. Анализ социологических данных. Методология, дескриптивная статистика, изучение связей между номинальными признаками. – М.: Научный мир, 2000. – 352 с.
21. Ферстер Э., Ренц Б. Методы корреляционного и регрессионного анализа. Руководство для экономистов: пер. с нем. – М.: Финансы и статистика, 1983. – 302 с.

Интернет-ресурсы

1. Бездудный Г.М., Знаменский В.А., Коваленко Н.В., Ковальчук В.Е., Луценко А.И., Рындина В.В. Задачи по математической статистике. Часть 2. Интервальное оценивание параметров распределения и критерии согласия. – Ростов-на-Дону: ЮФУ, 2007. – 34 с. URL: <http://window.edu.ru/resource/878/68878>
2. Блатов И.А., Старожилова О.В. Теория вероятностей и математическая статистика. Конспект лекций. – Самара: ПГУТИ, 2010. – 286 с. URL: <http://window.edu.ru/resource/896/74896/files/uch-pos-tv.pdf>
3. Борисова Е.В. Формирование и математическая обработка данных в социологии: учебное пособие. – Тверь:ТГТУ, 2006. – 120 с. URL: <http://window.edu.ru/resource/621/58621>
4. Васько О.Н., Капустин Е.И. Теория вероятностей и математическая статистика: опорный конспект и сборник задач. – Невинномысск: Невинномысский химический колледж, 2005. – 24 с. URL: <http://window.edu.ru/resource/945/53945/files/tv.pdf>
5. Воскобойников Ю.Е., Тимошенко Е.И. Математическая статистика: учебное пособие. – Новосибирск: Изд. НГАСУ, 2000. – 116 с. URL: <http://window.edu.ru/resource/304/63304>
6. Воскобойников Ю.Е., Тимошенко Е.И. Математическая статистика с примерами в Excel: учебное пособие. – Новосибирск: Изд. НГАСУ, 2006. – 154 с. URL: <http://window.edu.ru/resource/305/63305>
7. Выск Н.Д. Теория вероятностей и математическая статистика: учебное пособие. – М.: МАТИ-РГТУ им. К.Э. Циолковского, 2011. – 168 с. URL: <http://window.edu.ru/resource/889/76889/files/tv2011.pdf>
8. Галанов Ю.И. Математическая статистика: учебное пособие. – Томск: Изд-во ТПУ, 2010. – 80 с. URL: <http://window.edu.ru/resource/678/74678>
9. Компьютеризованный статистический анализ для историков. Учебное пособие. URL: <http://www.hist.msu.ru/Departments/Inf/Stud/textbook.htm>

10. А.И. Орлов Математика случая Вероятность и статистика – основные факты. Учебное пособие. – М.: МЗ-Пресс, 2004. URL: <http://www.aup.ru/books/m155/>
11. Палий И.А. Прикладная статистика: учебное пособие. – Омск: Изд-во СибАДИ, 2000. – Ч.1. – 79 с. URL: <http://window.edu.ru/resource/838/60838>
12. Романовский Р.К., Романовская А.М. Элементы теории вероятностей и математической статистики (теория и задачи): учебное пособие. – Омск: издатель ИП Скорнякова Е.В., 2012. – 189 с. URL: <http://window.edu.ru/resource/991/79991/files/romanovsky.pdf>
13. Салмин А.А. Анализ данных. Конспект лекций. – Самара.: ПГУТИ, 2013. – 111 с. URL: http://window.edu.ru/resource/815/78815/files/%D0%A1%D0%B0%D0%BB%D0%BC%D0%B8%D0%BD_%D0%90%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7%20%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85_%D0%9A%D0%BE%D0%BD%D1%81%D0%BF%D0%B5%D0%BA%D1%82%20%D0%BB%D0%B5%D0%BA%D1%86%D0%B8%D0%B9.pdf
- 14.** Статистика: Корреляционно-регрессионный анализ статистических связей на персональном компьютере: методические указания к практическим занятиям для студентов всех форм обучения специальности «Менеджмент» / Калинингр. ун-т; сост. Н.Ю. Лукьянова. – Калининград, 1999. URL: <http://www.aup.ru/files/m1021/m1021.pdf>

Интернет-ресурсы, содержащие статистическую информацию

Организация	Адрес сайта
Госкомстат РФ	http://www.gks.ru
Федеральная служба государственной статистики	http://www.fsgs.ru
Татарстанстат – Территориальный орган Федеральной службы государственной статистики по Республике Татарстан	http://www.tatstat.ru
Портал правительства республики Татарстан	http://prav.tatar.ru
Главный межрегиональный центр Росстата (Статистические организации)	http://www.gmcgks.ru/
Статистика российского образования	http://stat.edu.ru
ВЦИОМ	http://www.wciom.ru
Рекрутинговый портал SuperJob.ru	http://www.superjob.ru/
Министерство финансов РФ	http://www.minfin.ru
Министерство РФ по налогам и сборам	http://www.nalog.ru
Центральный банк РФ	http://www.cbr.ru
Экономика и жизнь: агентство консультаций и деловой информации	http://www.akdi.ru

ОТВЕТЫ

§2.1

3. 41-й размер, 23,26%.

5.

X	6	7	8	9	10	11	12	13	14
n_i^*	2	6	12	16	24	34	41	48	54

8.

X	17	18	19	20	21	22
n_i	3	5	1	3	4	3

§2.2

1. а) 5,86; б) 6; в) две моды – 3 и 6. $R = 9$; $S^2 = 8,4$. 2. а) 3,714; б) 3; в) 2. $S = 2,3$. 3. $Me=2$; медианным является домохозяйство, состоящее из двух человек; $\hat{Mo} = 1$; 2. 4. а) 32,4; 28,5; б) 22,8; 26,8. 5. $n = 56$; $\bar{x} = 3,66$; $C_v = 30,22$; $\hat{Mo} = 4$. 6. 2,14. 7. 3,47. 8. 1. а) 24,7; б) 24,75; в) 24,51; 2. 25,5.

9. Ответы вычислены с учетом того, что для интервала «Свыше 15 часов» в качестве верхней границы взято 25 часов.)

Показатели	Возрастные группы, лет		
	15–19	20–24	25–29
Среднее	5,26	4,15	3,4
Стандартное отклонение	3,98	3,97	3,24

10. 24,08 д.е.; $R = 9$ д.е. **11.** а) 6,01; б) 6,25; в) 6,09. **12.** 1) 236,11; 127,26. 2) 236,58; 120,15. **13.** 1) а) 59,13; б) 49,13; в) 51,9; г) 53,25. 2) 13,2. 3) а) 189,2; б) 152. **14.** 55,55%. **15.** а) 14,5; г) 2150; д) 102,38 . б) ; в)

Время опоздания, мин	Среднее время опозданий в каждой группе, мин	Общая сумма штрафов по каждой группе, д.е.
0–4	2,67	–
5–10	6,89	450
11–20	15,14	700
21–30	24,6	1000
Более 30	42,67	Выговор

16. 2) $\bar{x} = 71,85$; $S^2 = 83,23$.

§3.2

1. а) (45,99; 52,4); б) (45,62; 115,34). **2.** (1489,04; 1510,96), (1484,44; 1515,56). **3.** (2,1; 7,88). **4.** а) (17,73%; 31,65%); б) (21,69%; 36,34%). **5. 1)** а) (59,55; 61,45); б) (71,04; 72,56); 2) а) (19,78; 22,83); б) (53,76; 58,63); 3) . а) (2,07; 4,6); б) (2,5; 5,01). **6.** а) расчеты выполнены для последнего интервала (300–600). (229,01; 265,19); б) (0,0407; 0,123). **7.** 30,1. **8.** 97. **9.** 8,77; 10,45; 13,74. **10.** а) (17,14; 17,8); б) (0,397; 0,586). **11.** 1) (1,87%; 8,94%); 2) (38,04%; 56,76%).

§4.1

1. $t = -2$; гипотеза о среднем значении принимается. Данные подтверждают инструкцию на уровне значимости 0,05.
2. $t = 1,917$ а) гипотеза о среднем значении принимается; б) гипотеза отклоняется.
3. $t = 0,826$; гипотеза о среднем значении принимается.
4. $t = -2,59$; гипотеза о среднем значении принимается. Нет необходимости пересмотра нормы.
5. $t = -2,406$; гипотеза о среднем значении отклоняется.
6. $t = 3,23$; гипотеза о среднем значении отклоняется.

§4.2

1. $|t| = 5,961$; гипотеза о равенстве средних отклоняется. Различие существенно.
2. $|t| = 2,3$; а) гипотеза отклоняется; б) гипотеза принимается.
3. $|t| = 4,25$; гипотеза о равенстве средних отклоняется. В понимании счастья у различных возрастных групп наблюдаются значимые различия.
4. $|t| = 1,46$; при $\alpha = 0,05$ гипотеза о равенстве средних принимается. Результаты исследования не подтверждают значимых различий в зарплатах.
5. $|t| = 2,65$; гипотеза о равенстве средних отклоняется. Превышение средненедельного объема продаж в районе А по сравнению с районом В является статистически значимым.

- 6.** $|t| = 2,2$; 1) При $\alpha = 0,05$ гипотеза о равенстве средних отклоняется. Результаты исследования подтверждают выводы НР-агентств. 2) При $\alpha = 0,01$ гипотеза принимается. Расхождения в продолжительности поиска работы у соискателей двух групп статистически незначимо.
- 7.** $|t| = 1,06$; гипотеза о равенстве средних принимается. Полученные не подтверждают того, что сорт В урожайнее сорта А.

§4.3

- 1.** $f = 1,18$; гипотеза о равенстве дисперсий принимается.
- 2.** $f = 2,9$; гипотеза о равенстве дисперсий а) отклоняется; б) принимается.
- 3.** $f = 1,96$; гипотеза о равенстве дисперсий а) принимается; б) принимается.
- 4.** $f = 2,15$; гипотеза о равенстве дисперсий принимается. Станки обладают одинаковой точностью.
- 5.** $f = 3,87$; гипотеза о равенстве дисперсий отклоняется. Режимы работ трактористов А и В существенно отличны.
- 6.** $f = 1,39$; гипотеза о равенстве дисперсий принимается, следовательно, подтверждается однородность групп по результатам контрольной работы.
- 7.** $f = 4,94$; гипотеза о равенстве дисперсий принимается. Следовательно, по вариативности результатов контрольной работы учащиеся обоих направлений специализации также не отличаются
- 8.** $f = 1,3$; гипотеза о равенстве дисперсий принимается.
- 9.** $f = 1,043$; гипотеза о равенстве дисперсий принимается. Таким образом, полученные данные не подтверждают предположения экспер-

тов. Следовательно, можно считать, что активы А и В не отличаются по признаку вариативности.

§4.4

1. $z = 2,66$; гипотеза о равенстве долей отклоняется при $\alpha = 0,01$.
2. $z = 2,85$; гипотеза о равенстве долей отклоняется.
3. $z = 1,26$; гипотеза о равенстве долей принимается при $\alpha = 0,05$.
4. $z = 1,9$; гипотеза о равенстве долей принимается при $\alpha = 0,01$.
5. $z = 1,28$; гипотеза о равенстве долей принимается.
6. $z = 1,77$; гипотеза о равенстве долей принимается при $\alpha = 0,06$.
7. $|z| = 2,64$; гипотеза о равенстве долей отклоняется.
8. $z = 1,61$; гипотеза о равенстве долей принимается при $\alpha = 0,05$.

§4.5

1. 1) $n_{22} = 31$ – число людей в возрасте 22–29 лет, которые любят свою работу без фанатизма. 2) $n_{.2} = 91$ – из всех респондентов, участвовавших в опросе, 91 человек были в возрасте 22–29 лет 3) $n_{2.} = 136$; 4) 368; 5) 77; 6) 162; 7) а) 40,85%; б) 22,8%; 8) $\chi^2 = 23,73$. На уровне значимости 0,05 гипотеза отклоняется, признаки зависимы.

На уровне значимости 0,01 гипотеза о независимости признаков принимается.

2. $\chi^2 = 9,39$; гипотеза о независимости признаков отклоняется.
3. $\chi^2 = 32,52$; гипотеза о независимости признаков отклоняется, следовательно, признаки зависимы.

4. $\chi^2 = 7,45$; гипотеза принимается, признаки независимы. Независимо от возраста опрошенные предпочитают бумажные варианты книг.

5. Для следующей группировки

Возраст	18-25	26-30	30-45	46-60	Старше 60
Количество человек	6	7	9	14	8

$\chi^2 = 4,4$; гипотеза о равномерном распределении принимается.

6. $\chi^2 = 11,34$; гипотеза о независимости признаков принимается.

7. $\chi^2 = 8,86$; гипотеза о независимости признаков принимается. Т.е. не выявлено влияние пола на активность занятия спортом.

8. а) $\chi^2 = 8,05$; гипотеза о равномерном распределении отклоняется;

б) $\chi^2 = 1,04$; гипотеза о равномерном распределении принимается.

9. Ответ: $\chi^2 = 44,66$; гипотеза о независимости признаков отклоняется, следовательно подтверждается связь между производительностью труда и себестоимостью молока.

10. $\chi^2 = 0,5$; гипотеза о равномерном распределении принимается при $\alpha = 0,05$. Покупатели не выразили явного предпочтения ни одному из образцов новой упаковки.

11. $\chi^2 = 100,3$; гипотеза о независимости признаков отклоняется.

12. $\chi^2 = 11,02$; гипотеза о равномерном распределении принимается при $\alpha = 0,01$.

§5.1.1

1. Связи нет. **2.** Прямая связь. **3.** Обратная связь. **4.** Прямая связь.

5. $r = 0,802$; $Y = 0,43X - 1,4$. **6.** $r = 0,998$; $Y = -0,9 + 4,5X$; 30,6 тыс.р.

7. $r = 0,99$; $Y = 0,42X + 0,3$. **8.** $r = 0,89$; $Y = 4,4X - 5,7$. **9.** На уровне

значимости $\alpha = 0,05$ не подтверждается существенность статистической связи ($t = 2,12$). **10.** $r = -0.7$. При уровне значимости $\alpha = 0,1$ подтверждается существенность статистической связи ($t = -2,96$).

$Y = -0,22X + 2.59$. **11.** На уровне значимости $0,05$ подтверждается достоверность связи ($t = 9,33$). $Y = 2X + 24$; 27. **12.** 1) $\bar{X} = 7,05$; $\bar{Y} = 67,2$; 2) $r = 0,81$; $Y = 0,12X - 0,26$. **14.** $Y = 1,41X + 0,98$; 15,08.

§5.1.2.

1.

X	3	4	5	6	9	14
R	1	2	3	4	5	6

2.

X	2	3	3	3	4	4	9	11	12
R	1	3	3	3	5,5	5,5	7	8	9

3.

№ студента в списке	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Баллы за семестр	26	13	32	48	27	35	9	13	25	38	45	13	42	30
Ранг	6	3	9	14	7	10	1	3	5	11	13	3	12	8

§§5.1.2.1, 5.1.2.2

1. $r_s = 0,8$; $\tau_k = 0,6$. **2.** $r_s = 0,607$; гипотеза о равенстве нулю коэффициента Спирмена принимается на уровне значимости $\alpha = 0,05$.

$\tau_k = 0,33$. $t = 1,04$; гипотеза о равенстве нулю коэффициента Кендалла принимается на уровне значимости $\alpha = 0,05$.

3. а) $r_s = 0,7$; $\tau_k = 0,6$; б) $r_s = 0,9$; $\tau_k = 0,8$; в) $r_s = 1$; $\tau_k = 1$.

4. а) $r = 0,61$. $t = 2,3$; гипотеза о равенстве нулю линейного коэффициента отклоняется на уровне значимости $\alpha = 0,05$.

б) $\tau_k = 0,6$. $t = 2,57$; гипотеза о равенстве нулю коэффициента Кендалла отклоняется на уровне значимости $\alpha = 0,05$.

5. а) $r_s = 0,95$. $t = 8,6$; гипотеза о равенстве нулю коэффициента Спирмена отклоняется на уровне значимости $\alpha = 0,05$.

б) $r_s = 0,898$. $t = 5,77$; гипотеза о равенстве нулю коэффициента Спирмена отклоняется на уровне значимости $\alpha = 0,05$.

6. $r_s = 0,573$; гипотеза о равенстве нулю коэффициента Спирмена принимается на уровне значимости $\alpha = 0,05$.

§5.1.2.3

1. 0,83. **2.** 0,795. **3.** 0,89. **4.** 0,926. **5.** 0,86. **6.** 0,968.

§§5.2, 5.3

1. $r_{pb} = 0,24$; $r_{rb} = 0,31$. **2.** $r_{pb} = 0,24$; $r_{rb} = 0,31$. **3.** $r_{pb} = 0,6$. **4.** $r_{pb} = 0,29$; $r_{rb} = 0,35$. **5.** $r_{pb} = 0,07$; $r_{rb} = 0,21$. **6.** $r_{pb} = 0,43$; $r_{rb} = 0,36$. **7.** $r_{rb} = 0,6$.

§5.4

1. 1) $\varphi^2 = 0,18$; $P = 0,39$; $P' = 0,45$; $T = 0,23$; $K = 0,24$.

2) $\varphi^2 = 0,14$; $P = 0,35$; $P' = 0,4$; $T = 0,21$; $K = 0,21$.

2. $\varphi^2 = 0,06$; $P = 0,24$; $P' = 0,293$; $T = 0,16$; $K = 0,17$.

3. Ответы к задачам §4.5. Нумерация задач соответствует §4.5.

1. $\varphi^2 = 0,06$; $P = 0,25$; $P' = 0,284$; $T = 0,14$; $K = 0,15$.
2. $\varphi^2 = 0,11$; $P = 0,31$; $P' = 0,44$; $T = 0,27$; $K = 0,33$.
3. $\varphi^2 = 0,15$; $P = 0,36$; $P' = 0,51$; $T = 0,388$; $K = 0,29$.
9. $\varphi^2 = 0,27$; $P = 0,46$; $P' = 0,563$; $T = 0,37$; $K = 0,37$.
11. $\varphi^2 = 0,44$; $P = 0,55$; $P' = 0,637$; $T = 0,38$; $K = 0,38$.

§5.5

1. $Q = -0,88$; $\Phi = -0,59$.
2. $Q = 0,6$; $\Phi = 0,34$.
3. $Q = 0,43$; $\Phi = 0,22$.
4. $Q = 0,49$; $\Phi = 0,26$.
5. 1) $x^2 = 38,55$; 2) $x^2 = 23,72$; $Q = 0,51$; $\Phi = 0,27$.
6. $Q = 0,41$; $\Phi = 0,21$.

	Мотивация	
Баллы по КР	внешняя (0–10)	внутренняя (11–20)
КР < 6	9	6
КР > 6	5	8

7. $Q = 0,13$; $\Phi = 0,06$.
8. $Q = 0,37$; $\Phi = 0,19$.
9. $Q = 0,4$; $\Phi = 0,2$.
10. $\Phi = 0,6$.

This document was created with Win2PDF available at <http://www.daneprairie.com>.
The unregistered version of Win2PDF is for evaluation or non-commercial use only.